Asymptotically Optimal Algorithms for Budgeted Multiple Play Bandits

Alexander Luedtke, Emilie Kaufmann, Antoine Chambaz

Asymptotically Optimal Algorithms for Budgeted Multiple Play Bandits

Alex Luedtke[1], Emilie Kaufmann[2] and Antoine Chambaz[3]
[1]University of Washington, Department of Statistics.
[2]CNRS & Univ. Lille, CRIStAL (UMR 9189), Inria Lille.
[3]Université Paris Descartes, Laboratoire MAP5.

Abstract

We study a generalization of the multi-armed bandit problem with multiple plays where there is a cost associated with pulling each arm and the agent has a budget at each time that dictates how much she can expect to spend. We derive an asymptotic regret lower bound for any uniformly efficient algorithm in our setting. We then study a variant of Thompson sampling for Bernoulli rewards and a variant of KL-UCB for both single-parameter exponential families and bounded, finitely supported rewards. We show these algorithms are asymptotically optimal, both in rate and leading problem-dependent constants, including in the thick margin setting where multiple arms fall on the decision boundary.

1 Introduction

In the classical multi-armed bandit problem, an agent is repeatedly confronted with a set of K probability distributions $\nu_1, \dots, \nu_K$ called *arms* and must at each round select one of the available arms to pull based on their knowledge from previous rounds of the game. Each played arm presents the agent with a reward drawn from the corresponding distribution, and the agent's objective is to maximize the expected sum of their rewards over time or, equivalently, to minimize the total regret (the expected reward of pulling the optimal arm at every time step minus the expected sum of the rewards corresponding to their selected actions). To play the game well, the agent must balance the need to gather new information about the reward distribution of each arm (exploration) with the need to take advantage of the information that they already have by pulling the arm for which they believe the reward will be the highest (exploitation).

The bandit problem first started receiving rigorous mathematical attention slightly under a century ago [Thompson, 1933]. This early work focused on Bernoulli rewards, that are relevant in the simplest modeling of a sequential clinical trial, and presented a Bayesian algorithm now known as Thompson sampling. Since that time, many authors have contributed to a deeper understanding of the multi-armed bandit problem, both with Bernoulli and other reward distributions and either from a Bayesian [Gittins, 1979] or frequentist [Robbins, 1952] perspective. Lai and Robbins [1985] established a lower bound on the (frequentist) regret of any algorithm that satisfies a general uniform efficiency condition. This lower bound provides a concise definition of asymptotic (regret) optimality for an algorithm: an algorithm is asymptotically optimal when it achieves this lower bound. Lai [1987] introduced what are known as upper confidence bound (UCB) procedures for deciding which arm to pull at a given time step. In short, these procedures compute a UCB for the expected reward of each arm at each time and pull the arm with the highest UCB. Many variants of UCB algorithms have been proposed since then (see the Introduction of Cappé et al., 2013a for a thorough review), with more explicit indices and/or finite-time regret guarantees. Among them the KL-UCB algorithm [Cappé et al., 2013a] is proved to be asymptotically optimal for rewards that belong to a one-parameter exponential family and finitely-supported rewards. Meanwhile, there has been a recent interest in the theoretical understanding of the previously discussed Thompson

sampling algorithm, whose first regret bound was obtained by Agrawal and Goyal [2011]. Since then, Thompson Sampling has been proved to be asymptotically optimal for Bernoulli rewards [Kaufmann et al., 2012b, Agrawal and Goyal, 2012] and for reward distributions belonging to univariate exponential families [Korda et al., 2013].

There has recently been a surge of interest in the multi-armed bandit problem, due to its applications to (online) sequential content recommendation. In this context each arm models the feedback of an agent to a specific item that can be displayed (e.g. an advertisement). In this framework, it might be relevant to display *several* items at a time, and some variants of the classical bandit problems that have been proposed in the literature may be considered. In the *multi-armed bandit with multiple plays*, $m \geq 1$ out of K arms are sampled at each round and all the associated rewards are observed by the agent, who receives their sum. Anantharam et al. [1987] present a regret lower bound for this problem, together with a (non-explicit) matching strategy. More explicit strategies can be obtained when viewing this problem as a particular instance of a *combinatorial bandit problem with semi-bandit feedback*. Combinatorial bandits, originally introduced by Cesa-Bianchi and Lugosi [2012] in a non-stochastic setting, present the agent with possibly structured subsets of arms at each round: once a subset is chosen, the agent receives the sum of their rewards. The semi-bandit feedback corresponds to the case where the agent is able to see the reward of each of the sampled arms [Audibert et al., 2011]. Several extensions of UCB procedures have been proposed for the combinatorial setting (see e.g. Chen et al. [2013], Combes et al. [2015b]), with logarithmic regret guarantees. However, existing regret upper bounds do not match the lower bound of Anantharam et al. [1987]. In particular, despite the strong practical performance of KL-UCB-based algorithms in some combinatorial settings (including multiple-plays), their asymptotic optimality has never been established. Extending the optimality result from the single-play setting has proven challenging, especially in settings where the optimal set of m arms in non-unique. Recently, Komiyama et al. [2015] proved the asymptotic optimality of Thompson sampling for multiple-play bandits with Bernoulli rewards in the case where the arm with the m^{th} largest mean is unique. An important consequence of the uniqueness of the m^{th} largest mean is that the optimal set of m arms is necessarily unique, which may not be plausible in practice.

In this paper, we extend the multiple plays model in two directions, incorporating a *budget constraint* and an *indifference point*. Given a known cost c_a associated with pulling each arm a, at each round a subset of arms $\hat{\mathcal{A}}(t)$ is selected, so that the expected cost of pulling the chosen arms is at most the budget B. More formally, letting $C(t) \equiv \sum_{a \in \hat{\mathcal{A}}(t)} c_a$, one requires $\mathbb{E}[C(t)] \leq B$, where the expectation over the random selection of the subset $\hat{\mathcal{A}}(t)$ is taken conditionally on past observations. The agent observes the reward associated to the selected arms and receives a total reward $R(t) = \sum_{a=1}^{K} Y_a(t) \mathbb{1}_{(a \in \hat{\mathcal{A}}(t))}$, where $Y_a(t)$ is drawn from ν_a. This reward is then compared to what she could have obtained, had she spent the same budget on some other activity, for which the expect reward per cost unit is $\rho \geq 0$ (that is, the agent may prefer to use that money for some purpose that has reward to cost ratio greater than ρ and is external to the bandit problem). We note that, for positive reward distributions, choosing $\rho = 0$ corresponds to taking an action at every round. The agent's gain at round t is thus defined as

$$G(t) = R(t) - \rho C(t) = \sum_{a \in \hat{\mathcal{A}}(t)} (Y_a(t) - c_a \rho) \,.$$

The goal of the agent is to devise a sequential subset selection strategy that maximizes the expected sum of her gains, up to some horizon T and for which the budget constraint $\mathbb{E}[C(t)] \leq B$ is satisfied at each round $t \leq T$. In particular, arm a is "worth" drawing (in the sense that it increases the expected gain) only if its average reward per cost unit, μ_a / c_a (where μ_a is the expectation of ν_a), is at least the indifference point ρ.

This new framework no longer requires the number of arm draws to be fixed. Rather, the number of arm draws is selected to exhaust the budget, which makes sense in several online marketing scenarios. One can imagine for example a company targeting a new market on which it is willing to spend a budget B per week. Each week, the company has to decide which products to advertise for, and the cost of the advertising campaign may vary. After each week, the income associated to each campaign a is measured and compared to the minimal income of ρc_a that can be obtained when targeting other (known) markets

or investing the money in some other well-understood venture. Another possible scenario is that the same item can be displayed on several marketplaces never explored before for different costs, and the seller has to sequentially choose the different places he wants to display the items on while keeping the total budget spend smaller than B and maintaining a profitability larger than what can be obtained on a reference market place with reward per cost unit ρ.

Our first contribution is to characterize the best attainable performance in terms of regret (with respect to the gain $G(t)$, not the total reward $R(t)$) in this multiple-play bandit scenario with cost constraints, thanks to a lower bound that generalizes that of Anantharam et al. [1987]. We then study natural extensions of two existing bandit algorithms (KL-UCB and Thompson sampling) to our setting. We prove both rate and problem-dependent leading constant optimality for KL-UCB and Thompson sampling. The most difficult part of the proof is to show that the optimal arms away from the margin are pulled in almost every round (specifically, they are pulled in all but a sub-logarithmic number of rounds). Komiyama et al. [2015] studied this problem for Thompson sampling in multiple-play bandits using an argument different than that used in this paper. We provide a novel proof technique that leverages the asymptotic lower bound on the number of draws of any suboptimal arm. While this lower bound on suboptimal arm draws is typically used to prove an asymptotic lower bound on the regret of any reasonable algorithm, we use it as a key ingredient for our proof of an asymptotically optimal *upper bound* on the regret of KL-UCB and Thompson sampling, i.e. to prove the asymptotic optimality of these two algorithms. Also, throughout the manuscript, we do not assume that the set of optimal arms is unique, unlike most of the existing work on (standard) multiple-play bandits.

The rest of the article is organized as follows. Section 2 outlines our problem of interest. Section 3 provides an asymptotic lower bound on the number of suboptimal arm draws and on the regret. Section 4 presents the two sampling algorithms we consider in this paper and theorems establishing their asymptotic optimality: KL-UCB (Section 4.1) and Thompson sampling (Section 4.2). Section 5 presents numerical experiments supporting our theoretical findings. Section 6 presents the proofs of our asymptotic optimality (rate and leading constant) results for KL-UCB and Thompson Sampling. Section 7 gives concluding remarks. Technical proofs are postponed to the Appendix.

2 Multiple plays bandit with cost constraint

We consider a finite collection of arms $a \in \{1, \ldots, K\}$, where each arm has real-valued marginal reward distribution ν_a whose mean we denote by both μ_a and $E(\nu_a)$. Each arm belongs to a (possibly nonparametric) class of distributions $\mathcal{D}$. We use $\mathcal{V}$ to denote $(\nu_1, \ldots, \nu_K)$, where $\mathcal{V}$ belongs to any model $\mathcal{D}_K$ that is variation-independent in the sense that, for each $a \in \{1, \ldots, K\}$, knowing the joint distribution of the rewards $a' \neq a$ places no restrictions on the collection of possible marginal distributions of ν_a, i.e. ν_a could be equal to any element in $\mathcal{D}$. More formally, letting $\mathcal{D}_{-a}$ denote the collection of joint distributions of the rewards $a' \neq a$ implied by at least one distribution in $\mathcal{D}_K$, variation independence states that, for each $a \in \{1, \ldots, K\}$ it is true that, for every joint distribution $V_{-a} \in \mathcal{D}_{-a}$ and every distribution $\nu_a \in \mathcal{D}$, there exists a distribution in $\mathcal{D}_K$ whose joint distribution of the rewards $a' \neq a$ is equal to V_{-a} and whose marginal distribution of reward a is equal to ν_a. An example of a statistical model satisfying this variation-independence assumption is the distribution in which the rewards of all of the arms are independent and the marginal distributions ν_a fall in $\mathcal{D}$ for all a, though this assumption also allows for high levels of dependence between the rewards of the arms, i.e. is not to be confused with the *much stronger* model assumption of independence between the different arms.

2.1 The sequential decision problem

Let $\{(Y_1(t), \ldots, Y_K(t))\}_{t=1}^{\infty}$ be an independent and identically distributed (i.i.d.) sample from the distribution $\mathcal{V}$. In the multiple-play bandit with cost constraint, each arm a is associated with a known *cost* $c_a > 0$. The model also depends on a known *budget per round* B and *indifference parameter* $\rho \geq 0$. At round t, the agent selects a subset $\hat{\mathcal{A}}(t)$ of arms and subsequently observes the action-reward pairs

$\{(a, Y_a(t)) : a \in \hat{\mathcal{A}}(t)\}$. We emphasize that the agent is aware that reward $Y_a(t)$ corresponds to the action $a \in \hat{\mathcal{A}}(t)$. This subset $\hat{\mathcal{A}}(t)$ is drawn from a distribution $Q(t-1)$ over $\mathcal{S}_K$, the set of all subsets of $\{1, \ldots, K\}$, that depends on the observations gathered at the $(t-1)$ previous rounds. More precisely, $Q(t)$ is $\mathcal{F}(t)$-measurable, where $\mathcal{F}(t)$ is the σ-field generated by all action-reward pairs seen at times $1, \ldots, t$, and possibly also some exogenous stochastic mechanism. We use $q_a(t)$ to denote the probability that arm a falls in $\hat{\mathcal{A}}(t+1) \sim Q(t)$.

Given the budget B and the indifference parameter ρ, at each round $(t+1)$ the distribution $Q(t)$ must respect the budget constraint

$$\mathbb{E}_{\mathcal{A} \sim Q(t)} \left[\sum_{a \in \mathcal{A}} c_a \right] \leq B, \quad \text{or, equivalently,} \quad \sum_{a=1}^{K} c_a q_a(t) \leq B. \tag{1}$$

Upon selecting the arms, the agent receives a reward $R(t+1) = \sum_{a \in \hat{\mathcal{A}}(t+1)} Y_a(t+1)$ and incurs a gain $G(t+1) = \sum_{a \in \hat{\mathcal{A}}(t+1)} (Y_a(t+1) - c_a \rho)$. Given a (possibly unknown) horizon T, the goal of the agent is to adopt a strategy for sequentially selecting the distributions $Q(t)$ that maximizes

$$\mathbb{E}\left[\sum_{t=1}^{T} G(t) \right],$$

while satisfying, at each round $t = 0, \ldots, T-1$ the budget constraint (1). This constraint may be viewed as a 'soft' budget constraint, as it allows the agent to (slightly) exceed the budget at some rounds, as long as the expected cost remains below B at each round. We shall see below that considering a 'hard' budget constraint, that is selecting at each round a deterministic subset $\hat{\mathcal{A}}(t)$ that satisfies $\sum_{a=1}^{K} c_a \mathbb{1}_{(a \in \hat{\mathcal{A}}(t))} \leq B$, is a much harder problem. Besides, in the marketing examples described in the introduction, it makes sense to consider a large time horizon and to allow for minor budget crossings. Under the soft budget constraint (1), if we knew the vector of expected mean rewards $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_K)$, at each round t we would draw a subset from a distribution

$$Q^\star \in \operatorname*{argmax}_{Q} \mathbb{E}_{S \sim Q} \left[\sum_{a \in S} (\mu_a - c_a \rho) \right] \quad \text{such that} \quad \mathbb{E}_{S \sim Q} \left[\sum_{a \in S} c_a \right] \leq B. \tag{2}$$

Above, the argmax is over distributions Q with support on the power set of $\{1, \ldots, K\}$. Noting that the two expectations only depend on the marginal probability of inclusions $q_a = \mathbb{P}_{S \sim Q}(a \in S)$, it boils down to finding a vector $\boldsymbol{q}^\star = (q_a)_{a=1}^K$ that satisfies

$$\boldsymbol{q}^\star \in \operatorname*{argmax}_{\boldsymbol{q} \in [0,1]^K} \sum_{a=1}^{K} q_a (\mu_a - c_a \rho) \quad \text{such that} \quad \sum_{a=1}^{K} q_a c_a \leq B. \tag{3}$$

An oracle strategy would then draw S from a distribution $Q^\star$ with marginal probabilities of inclusions given by $\boldsymbol{q}^\star$ (e.g. including independently each arm a with probability $q_a^\star$). The optimization problem (3) is known as a fractional knapsack problem [Dantzig, 1957], and its solution is a greedy strategy, that is described below. It is expressed in terms of the reward-to-cost ratio of each arm a, defined as $\rho_a \equiv \mu_a / c_a$.

Proposition 1. *Introduce*

$$\rho^\star \equiv \begin{cases} \rho \quad \textit{if} \quad \sum_{a:\rho_a > \rho} c_a < B, \\ \sup\{r \geq 0 : \sum_{a:\rho_a > r} c_a \geq B\} \geq \rho \quad \textit{otherwise}, \end{cases}$$

and define the three sets

$$\begin{aligned} \textit{optimal arms away from the margin:} \quad & \mathcal{L} \equiv \{a : \rho_a > \rho^\star\}, \\ \textit{arms on the margin:} \quad & \mathcal{M} \equiv \{a : \rho_a = \rho^\star\}, \end{aligned}$$

suboptimal arms away from the margin: $\mathcal{N} \equiv \{a : \rho_a < \rho^\star\}$.

Then $\boldsymbol{q}^\star$ *is solution to* (3) *if and only if* $q_a^\star = 1$ *for all* $a \in \mathcal{L}$, $q_b^\star = 0$ *for all* $b \in \mathcal{N}$ *and* $\sum_{a\in\mathcal{M}} c_a q_a^\star = B - \sum_{a\in\mathcal{L}} c_a$ *if* $\rho^\star > \rho$.

We would like to emphasize that, just like the quantities $Q^\star$, $q^\star$ or ρ_a defined above, the quantity $\rho^\star$ defined in Proposition 1 depends on the value of ρ, the vector of cost and on the vector of means $\boldsymbol{\mu}$. When we need to materialize this dependency in $\boldsymbol{\mu}$ we shall use the notation $\rho^\star(\boldsymbol{\mu})$, but it is sometimes omitted for the sake of readability.

From Proposition 1, proved in Appendix A, the optimal strategy sorts the items by decreasing order of ρ_a, and includes them one by one ($q_a^\star = 1$), as long as the value increases and the budget is not exceeded. Then we can identify two situations: if $\rho^\star(\boldsymbol{\mu}) = \rho$, there are not enough interesting items (i.e. such that $\rho_a > \rho$) to saturate the budget, and the optimal strategy is to include all the interesting items. If $\rho^\star(\boldsymbol{\mu}) > \rho$, some probability of inclusion is further given to the items on the margin in order to saturate the budget constraint. In that case, the margin is always non-empty: there exist items a such that $\rho^\star(\boldsymbol{\mu}) = \rho_a$.

Recovering the multiple-play bandit model. By choosing $c_a = 1$ for all arm a, $B = m$ and $\rho = 0$, we recover the classical multiple-play bandit model. In that case $\rho^\star(\boldsymbol{\mu}) = \mu_{[m]}$, where $[m]$ is the arm with the m^{th} largest mean and $Q^\star = \delta_{\{[1],\dots,[m]\}}$ is a solution to (2): the corresponding oracle strategy always plays the m arms with largest means.

Hard and soft constraints. Under hard budget constraints, if we knew the vector of expected mean rewards $\boldsymbol{\mu}$, at each round t we would pick a subset

$$S^\star \in \operatorname*{argmax}_{S\in\mathcal{S}_K} \sum_{a\in S}(\mu_a - c_a\rho) \quad \text{such that} \quad \sum_{a\in S} c_a \leq B. \tag{4}$$

This is a 0/1 knapsack problem, that is much harder to solve than the above fractional knapsack problem. In fact, 0/1 knapsack problems are NP-hard, though they are, admittedly, some of the easiest problems in this class, and reasonable approximation schemes exist [Karp, 1972]. Nonetheless, the greedy strategy (including arms by decreasing order of ρ_a while the budget is not exceeded, with ties broken arbitrarily) is not generally a solution to (4). However, using Proposition 1, one can identify some examples where there exist deterministic solutions to (3), i.e. solutions such that $q_a^\star \in \{0, 1\}$ that are therefore solutions to (4): if $\rho^\star(\boldsymbol{\mu}) = \rho$ or if there exists $m \in \mathcal{M}$ such that $\sum_{a\in\mathcal{L}\cup\{m\}} c_a = B$. Hence the multiple-play bandit model can be viewed as a particular instance of the multiple plays model under both hard or soft budget constraint. In the rest of the article, we only consider soft budget constraints, as there is generally no tractable oracle under hard budget constraints.

High-probability bound on the budget spent by a finite horizon T. In Appendix **??**, we outline how one could analyze the regret of algorithms that respect the soft budget constraint (1) at each time t in a finite-horizon problem in which the requirement that (1) hold at each time t is replaced by the hard budget constraint that $\sum_{t=1}^T \sum_{a\in\hat{\mathcal{A}}(t)} c_a \leq BT$ almost surely. Our argument suggests that the regret in these settings should be no worse than $O(\sqrt{T})$.

2.2 Regret decompositions

The best achievable (oracle) performance consists in choosing, at every round t, $Q(t)$ to be the optimal distribution $Q^\star$ whose probabilities of inclusions are described in Proposition 1. Using the definitions introduced in Proposition 1, such a strategy ensures an expected gain at each round of

$$G^\star \equiv \sum_{a=1}^K q_a^\star(\mu_a - c_a\rho). \tag{5}$$

The quantity above is the reward from pulling the chosen arms relative to the reward from reallocating the expected cost of the strategy, namely $\sum_{a=1}^{K} q_a^\star c_a$, to pursue the action (which is external to the bandit problem) that has reward-to-cost ratio equal to the indifference point ρ. We prove the following identity in Appendix A.

Proposition 2. *It holds that*

$$G^\star = \sum_{a\in\mathcal{L}} \mu_a + \rho^\star \left(B - \sum_{a\in\mathcal{L}} c_a \right) - B\rho.$$

Maximizing the expected total gain is equivalent to minimizing the regret, that is the difference in performance compared to the oracle strategy:

$$\text{Regret}(T, \mathcal{V}, \texttt{Alg}) \equiv TG^\star - \mathbb{E}_{\mathcal{V}} \left[\sum_{t=1}^{T} G(t) \right],$$

where the sequence of gains $G(t)$ is obtained under algorithm `Alg`. The following statement, proved in Appendix A, provides an interesting decomposition of the regret, as a function of the number of selections of each arm, denoted by $N_a(T) \equiv \sum_{t=1}^{T} \mathbb{1}\{a \in \hat{\mathcal{A}}(t)\}$.

Proposition 3. *With $\rho^\star = \rho^\star(\boldsymbol{\mu})$, $\mathcal{L}, \mathcal{N}$ defined as in Proposition 1, for any algorithm `Alg`*

$$\begin{aligned} \text{Regret}(T, \mathcal{V}, \texttt{Alg}) \quad = \quad & \sum_{a^\star\in\mathcal{L}} c_{a^\star} (\rho_{a^\star} - \rho^\star) \left(T - \mathbb{E}_{\mathcal{V}}[N_{a^\star}(T)]\right) + \sum_{a\in\mathcal{N}} c_a [\rho^\star - \rho_a] \, \mathbb{E}_{\mathcal{V}}[N_a(T)] \\ & + (\rho^\star - \rho) \left(BT - \sum_{a=1}^{K} c_a \mathbb{E}_{\mathcal{V}}[N_a(T)] \right). \end{aligned} \tag{6}$$

This decomposition writes the regret as a sum of three non-negative terms. In order for the regret to be small, each optimal arm $a^\star \in \mathcal{L}$ should be drawn very often (of order T times, to make the first term small) and each suboptimal arm $a^\star \in \mathcal{N}$ should be drawn seldomly (to make the second term small). Finally if $\rho^\star > \rho$, that is if there are sufficiently many 'worthwhile' arms to exceed the budget, then the third term appears as a penalty for not using the whole budget at every round. It means that arms on the margin $\mathcal{M}$ have to be drawn sufficiently often so as to saturate the budget constraint.

An extended bandit interpretation. Here we propose another view on this regret decomposition, by means of an extended bandit game with an extra arm, which we term a pseudo-arm, that represents the choice not to pull arms. Whenever an algorithm does not saturate the budget constraint (1), one can view this algorithm as putting weight on a pseudo-arm in the bandit, that yields zero gain but permits saturation of the budget. Letting $\mu_{K+1} = B\rho$ and $c_{K+1} = B$, the gain associated with drawing arm $(K+1)$ (whose distribution is a point mass at $B\rho$) is indeed zero (as $\mu_{K+1} - \rho c_{K+1} = 0$) and, for any $\boldsymbol{q}(t)$ such that $\sum_{a=1}^{K} q_a(t) c_a \leq B$, there exists $q_{K+1}(t)$ such that $\sum_{a=1}^{K+1} q_a(t) c_a = B$, as $c_{K+1} = B$. Any algorithm for the original bandit problem selecting $\hat{S}(t) \in \mathcal{S}_K$ at time t can thus be viewed as an algorithm selecting $\tilde{S}(t) \in \mathcal{S}_{K+1}$, that additionally includes arm $(K+1)$ with probability $q_{K+1}(t)$. As the pseudo-arm is associated with a null gain, the cumulated gain and regret are similar in both settings. Moreover, as $q_{K+1}(t) = (B - \sum_{a=1}^{K} c_a q_a(t))/B$, one easily sees that the number of (artificial) selections of the pseudo-arm is such that

$$B\mathbb{E}[N_{K+1}(T)] = BT - \sum_{a=1}^{K} c_a \mathbb{E}[N_a(T)],$$

which equals the third term in the regret decomposition, up to the factor $(\rho^\star - \rho)$.

In this extended bandit model, the three sets of arms introduced in Proposition 1 remain unchanged, with $\mathcal{L} \equiv \{a \in \{1,\dots,K+1\} : \rho_a > \rho^\star\}$, $\mathcal{M} \equiv \{a \in \{1,\dots,K+1\} : \rho_a = \rho^\star\}$ and $\mathcal{N} \equiv \{a \in \{1,\dots,K+1\} : \rho_a < \rho^\star\}$. As $\rho_{K+1} = \rho \leq \rho^\star$, the pseudo-arm may only belong to $\mathcal{M}$ or $\mathcal{N}$, and the margin $\mathcal{M}$ is always non-empty. Considering the extended bandit model, the regret decomposition can be rewritten in a more compact way:

$$\mathrm{Regret}(T, \mathcal{V}, \mathtt{Alg}) = \sum_{a^\star \in \mathcal{L}} c_{a^\star} (\rho_{a^\star} - \rho^\star) (T - \mathbb{E}[N_{a^\star}(T)]) + \sum_{a \in \mathcal{N}} c_a [\rho^\star - \rho_a] \, \mathbb{E}[N_a(T)].$$

Our proofs make use of this extended bandit model, since many of the results we present apply to both the "actual" arms $a = 1,\dots,K$ and the pseudo-arm $(K+1)$. Our proofs also make use of a set $\mathcal{S}$, which, in the extended bandit model, refers to all arms in $(\mathcal{L} \cup \mathcal{M}) \backslash \{K+1\}$ whereas, in the unextended bandit model, it refers simply to all optimal arms both on and away from the margin.

2.3 Related work

There has been considerable work on various forms of "budgeted" or "knapsack" bandit problems [Tran-Thanh et al., 2012, Badanidiyuru et al., 2013, Agrawal and Devanur, 2014, Xia et al., 2015, 2016a, Li and Xia, 2017]. The main difference between our work and these works is that we consider a round-wise budget constrain, and allow for several arms to be selected at each round, possibly in a randomized way in order to satisfy the budget constraint in expectation. In contrast, in most existing works, one arm is (deterministically) selected at each round, and the game ends when a global budget is exhausted. The work of Xia et al. [2016b] appears to be the most closely related to ours: in their setup the agent may play multiple arms at each round, though the number of arms pulled at each round is fixed and the cost of pulling each arm is random and observed upon pulling each arm. Sankararaman and Slivkins [2018] also consider a framework in which a subset of arms is selected at each round, but this subset is chosen from a list of candidate subsets (as in a combinatorial bandit problem) and there is a global budget constraint. Compared to all these mentioned budgeted bandit problems, the focus of our analysis differs substantially, in that our primary objective is to not only prove rate optimality, but also leading constant optimality of our regret bounds. Proving constant optimality is especially challenging in situations where the set of optimal arms is non-unique, but we give careful arguments that overcome this challenge.

Several other extensions of the multiple-play bandit model have been studied in the literature. UCB algorithms have been widely used in the combinatorial semi-bandit setting, in which at each time step a subset of arms has to be select among a given class of subsets, and the rewards of every individual arms in the subset are observed. The most natural use of UCBs and the "optimism in face of uncertainty principle" is to choose at every time step the subset that would be the best if the unknown means were equal to the corresponding UCBs. This was studied by Chen et al. [2013], Kveton et al. [2014], Wen et al. [2015], who exhibit good empirical performance and logarithmic regret bounds. Combes et al. [2015b] further study instance-dependent optimality for combinatorial semi bandits, and propose an algorithm based on confidence bounds on the value of each subset, rather than on confidence bounds on the arms' means. Their ESCB algorithm is proved to be order-optimal for several combinatorial problems. As a by product of our results, we will see that in the multiple-play setting, using KL-based confidence bounds on the arms' means is sufficient to achieve asymptotic optimality. Another interesting direction of extension is the possibility to have only partial feedback over the m proposed item. Variants of KL-UCB and Thompson Sampling were proposed for the Cascading bandit model [Kveton et al., 2015a,b], Learning to Rank [Combes et al., 2015a] or the Position-Based model [Lagrée et al., 2016]. It would be interesting to try to extend the results presented in this work to these partial feedback settings.

3 Regret Lower Bound

We first give in Lemma 4 asymptotic lower bounds on the number of draws of suboptimal arms, either in high-probability or in expectation, in the spirit of those obtained by Lai and Robbins [1985], Anantharam

et al. [1987]. Compared to these works, the lower bounds obtained here hold under our more general assumptions on the arm distributions, which is reminiscent of the work of Burnetas and Katehakis [1996].

To be able to state our regret lower bound, we now introduce the following notation. We let $\mathrm{KL}(\nu, \nu')$ denote the KL-divergence between distributions ν and ν'. If ν and ν' are uniquely parameterized by their respective means μ and μ' as in a canonical single parameter exponential family (e.g. Bernoulli distributions), then we abuse notation and let $\mathrm{KL}(\mu, \mu') \equiv \mathrm{KL}(\nu, \nu')$. For a distribution $\nu \in \mathcal{D}$ and a real μ, we define

$$\mathcal{K}_{\inf}(\nu, \mu) \equiv \inf \left\{ \mathrm{KL}(\nu, \nu') : \nu' \in \mathcal{D} \text{ and } \mu < E(\nu') \text{ and } \nu \ll \nu' \right\}, \tag{7}$$

with the convention that $\mathcal{K}_{\inf}(\nu, \mu) = \infty$ if there does not exist a $\nu \ll \nu'$ with $\mu < E(\nu')$. We will also use the convention that, for finite constants d_1, d_2, $d_1/(d_2 + \mathcal{K}_{\inf}(\nu, \mu)) = 0$ when $\mathcal{K}_{\inf}(\nu, \mu) = \infty$. We make one final assumption, and introduce two disjoint sets $\underline{\mathcal{N}}$ and $\overline{\mathcal{N}}$, whose union is $\mathcal{N}$. The assumption is that, for each arm $a \in \{1, \ldots, K\}$, μ_a falls below the upper bound of the expected reward parameter space, i.e. $\mu_a < \mu_+ \equiv \sup\{E(\nu) : \nu \in \mathcal{D}\}$. We define the sets $\underline{\mathcal{N}}$ and $\overline{\mathcal{N}}$ respectively as the subsets of $\mathcal{N}$ for which optimality is and is not feasible given our parameter space, namely

$$\underline{\mathcal{N}} \equiv [\mathcal{N} \cap \{a : c_a \rho^\star < \mu_+\}] \setminus \{K+1\}$$
$$\overline{\mathcal{N}} \equiv [\mathcal{N} \cap \{a : c_a \rho^\star \geq \mu_+\}] \setminus \{K+1\}.$$

By defining $\underline{\mathcal{N}}$ and $\overline{\mathcal{N}}$ in this way, these sets agree in the extended and unextended bandit models. The lower bounds presented in this section will also agree in these two models.

We now define a uniformly efficient algorithm, that generalizes the class of algorithms considered in Lai and Robbins [1985]. An algorithm `Alg` is uniformly efficient if, for all $\mathcal{V} \in \mathcal{D}_K$ and $\alpha \in (0, 1)$, $\mathrm{Regret}(T, \mathcal{V}, \texttt{Alg}) = o(T^\alpha)$ as T goes to infinity (from now on, the limits in T will be for $T \to \infty$). From the regret decomposition (6), this is equivalent to

1. $T - \mathbb{E}_{\mathcal{V}}\left[N_{a^\star}(T)\right] = o\left(T^\alpha\right)$ for all arms $a^\star$ such that $\rho_{a^\star} > \rho^\star(\boldsymbol{\mu})$;
2. $\mathbb{E}_{\mathcal{V}}[N_a(T)] = o(T^\alpha)$ for all arms a such that $\rho_a < \rho^\star(\boldsymbol{\mu})$;
3. if $\rho^\star(\boldsymbol{\mu}) > \rho$, $BT - \sum_{a=1}^K c_a \mathbb{E}_{\mathcal{V}}[N_a(T)] = o(T^\alpha)$,

where above and throughout we write $\mathbb{E}_{\mathcal{V}}$ when we wish to emphasize that the expectation is over $\mathcal{V}$.

Lemma 4 (Lower bound on suboptimal arm draws)**.** *If an algorithm is uniformly efficient, then, for any arm $a \in (\mathcal{M} \cup \underline{\mathcal{N}}) \setminus \{K+1\}$ and any $\delta \in (0, 1)$ and $\epsilon > 0$,*

$$\lim_T \mathbb{P}\left\{ N_a(T) < (1-\delta) \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, c_a \rho^\star) + \epsilon} \right\} = 0. \tag{8}$$

One can take $\epsilon = 0$ if $a \in \underline{\mathcal{N}}$. Furthermore, for any suboptimal arm $a \in \underline{\mathcal{N}}$,

$$\liminf_T \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{\mathcal{K}_{\inf}(\nu_a, c_a \rho^\star)}. \tag{9}$$

We defer the proof of this result to Appendix B. We note that, while (9) could also easily be obtained using the recent change-of-distribution tools introduced by Garivier et al. [2016], we need to go back to Lai and Robbins' technique to prove the high-probability result (8), which will be crucial in the sequel. Indeed, we will use it to prove optimal regret of our algorithms: in essence we need to ensure that we have enough information about arms in $\mathcal{M} \cup \mathcal{N}$ to ensure that we pull the optimal arms in $\mathcal{L}$ sufficiently often.

We now present a corollary to Lemma 4 which provides a regret lower bound, as well as sufficient conditions for an algorithm to asymptotically match it. As already noted by Komiyama et al. [2015] in

the Bernoulli case for the bandit with multiple-play problems, an algorithm achieving the asymptotic lower bound (9) on the expected number of draws of arms in $\underline{\mathcal{N}}$ does not necessarily achieve optimal regret, unlike in classic bandit problems. Thus, we emphasize that the upcoming condition (11) alone is not sufficient to prove asymptotic optimality. The conditions of this proof can be easily obtained from the regret decomposition (6), and so the proof is omitted.

Theorem 5 (Regret lower bound)**.** *If an algorithm* `Alg` *is uniformly efficient, then*

$$\liminf_T \frac{\mathrm{Regret}(T, \mathcal{V}, \texttt{Alg})}{\log T} \geq \sum_{a \in \underline{\mathcal{N}}} \frac{c_a(\rho^\star - \rho_a)}{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)}. \tag{10}$$

Moreover, any algorithm `Alg` *satisfying*

$$\text{for arms } a \in \underline{\mathcal{N}}: \quad \mathbb{E}_{\mathcal{V}}[N_a(T)] = \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)} + o(\log T), \tag{11}$$

$$\text{for arms } a \in \overline{\mathcal{N}} \quad \mathbb{E}_{\mathcal{V}}[N_a(T)] = o(\log T), \tag{12}$$

$$\text{for arms } a^\star \in \mathcal{L}: \quad \mathbb{E}_{\mathcal{V}}[N_{a^\star}(t)] = T - o(\log T), \tag{13}$$

and, if $\rho^\star(\boldsymbol{\mu}) > \rho$,

$$BT - \sum_{a=1}^{K} c_a \mathbb{E}_{\mathcal{V}}[N_a(T)] = o(\log(T)), \tag{14}$$

is asymptotically optimal, in the sense that it satisfies

$$\limsup_T \frac{\mathrm{Regret}(T, \mathcal{V}, \texttt{Alg})}{\log T} \leq \sum_{a \in \underline{\mathcal{N}}} \frac{c_a(\rho^\star - \rho_a)}{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)}. \tag{15}$$

4 Algorithms

Algorithms rely on estimates of the arm distributions and their means, that we formally introduce below. For each arm a and natural number n, define $\tau_{a,n} = \min\{t \geq 1 : N_a(t) = n\}$ to be the (stopping) time at which the n^{th} draw of arm a occurs. Let $X_{a,n} \equiv Y_a(\tau_{a,n})$ denote the n^{th} draw from ν_a. One can show that $\{X_{a,n}\}_{n=1}^{\infty}$ is an i.i.d. sequence of draws from ν_a for each a, though we note that our variation independence assumption is too weak to ensure that these sequences are independent for two arms $a \neq a'$ (this is not problematic – most of our arguments end up focusing on arm-specific sequences $\{X_{a,n}\}_{n=1}^{\infty}$)[1]. We denote the empirical distribution function of observations drawn from arm a by any time T by

$$\hat{\nu}_a(T) \equiv \frac{1}{N_a(T)} \sum_{t=1}^{T} \delta_{Y_a(t)} \mathbb{1}\{a \in \hat{\mathcal{A}}(t)\} = \frac{1}{N_a(T)} \sum_{n=1}^{N_a(T)} \delta_{X_{a,n}}.$$

We similarly define $\hat{\nu}_{a,n}$ to be the empirical distribution function of the observations $X_{a,1}$, …, $X_{a,n}$. Thus, $\hat{\nu}_a(t) = \hat{\nu}_{a,N_a(t)}$. We further define $\hat{\mu}_a(t)$ to be the empirical mean of observations drawn from arm a by time t and $\hat{\mu}_{a,N_a(t)} = \hat{\mu}_a(t)$.

4.1 KL-UCB

At time t, UCB algorithms leverage high probability upper bound $U_a(t)$ on μ_a for each a. The methods used to build these confidence bounds vary, as does the way the algorithm uses these confidence

[1]It is *a priori* possible that $\tau_{a,n} = \infty$ for all n large enough (though, as we showed in Section 3, this event will occur with probability zero for any reasonable algorithm). To deal with this case, let $X_{a,n} \equiv Y_a(\tau_{a,n})$ denote the n^{th} draws from ν_a for all $\tau_{a,n} < \infty$ and let $\{X_{a,n}\}_{n:\tau_{a,n}=\infty}$ denote an i.i.d. sequence independent of $\{X_{a,n}\}_{n:\tau_{a,n}<\infty}$.

Algorithm KL-UCB

Parameters A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$ and an operator $\Pi_{\mathcal{D}}$ mapping each empirical distribution functions $\hat{\nu}_a(t)$ to an element of the model $\mathcal{D}$.
Initialization Pull each arm of $\{1, \ldots, K\}$ once.[2]
for $t = K, K+1, \ldots, T-1$ **do**
  For $a = 1, \ldots, K$, let $U_a(t)$ be defined as in (16).
  Let $\hat{\rho}^\star(t) \equiv \rho^\star\left(U_a(t) : a = 1, \ldots, K\right)$.
  For $a \in \{1, \ldots, K\}$, let $q_a(t) = \mathbb{1}\{U_a(t) > c_a \hat{\rho}^\star(t)\}$.
  if $\widehat{\mathcal{M}}(t) \equiv \{a : U_a(t) = c_a \hat{\rho}^\star(t)\}$ is non-empty **then**
    if $\hat{\rho}^\star(t) > \rho$ **then**
      For $a \in \widehat{\mathcal{M}}(t)$, let $q_a(t) = \left[B - \sum_{a:U_a(t) > c_a \hat{\rho}^\star(t)} c_a\right] / \sum_{a \in \widehat{\mathcal{M}}(t)} c_a$.
    else Let $q_a(t) = 0$ for all $a \in \widehat{\mathcal{M}}(t)$[3].
  Draw $\hat{\mathcal{A}}(t+1)$ from any distribution $Q(t)$ with marginal probabilities $q_a(t)$[4].
  Draw arms in $\hat{\mathcal{A}}(t+1)$ and observe $Y_a(t+1)$, $a \in \hat{\mathcal{A}}(t+1)$.

bounds. In our setting, we derive these bounds using the same technique as for KL-UCB in Cappé et al. [2013a]. At the beginning of round $(t+1)$, the KL-UCB algorithm computes an optimistic oracle strategy $(q_a(t))_{a=1,\ldots,K}$, that is an oracle strategy assuming the unknown mean of each arm a is equal to its best possible value, $U_a(t)$. From Proposition 1, this optimistic oracle depends on $\hat{\rho}^\star(t) = \rho^\star\left(U_a(t) : a = 1, \ldots, K\right)$, where $\rho^\star(\boldsymbol{\mu})$ is the function defined in Proposition 1. Then each arm is included in $\hat{\mathcal{A}}(t+1)$ independently with probability $q_a(t)$. Due to the structure of an oracle strategy, KL-UCB can be rephrased as successively drawing the arms by decreasing order of the ratio $U_a(t)/c_a$ until the point that the budget is exhausted, with some probability to include the arms on the margin. We choose to keep the name KL-UCB for this straightforward generalization of the original KL-UCB algorithm.

The definition of the upper bound $U_a(t)$ is closely related to that of $\mathcal{K}_{\inf}$ given in (7). Let $\Pi_{\mathcal{D}}$ be a problem-specific operator mapping each empirical distribution function $\hat{\nu}_a(t)$ to an element of the model $\mathcal{D}$. Furthermore, let $f : \mathbb{N} \to \mathbb{R}$ be a non-decreasing function, where this function is usually chosen so that $f(t) \approx \log t$. The UCB is then defined as

$$U_a(t) \equiv \sup \left\{ E(\nu) : \nu \in \mathcal{D} \text{ and } \mathrm{KL}\left(\Pi_{\mathcal{D}}\left(\hat{\nu}_a(t)\right), \nu\right) \leq \frac{f(t)}{N_a(t)} \right\}, \; a = 1, \ldots, K. \tag{16}$$

As we will see, the closed form expression for $U_a(t)$ can be made slightly more explicit for exponential family models, though the expression still has the same general flavor. If a number μ satisfies $\mu \geq U_a(t)$, then this implies that, for every $\nu \in \mathcal{D}$ for which $E(\nu) > \mu$, $\mathrm{KL}\left(\Pi_{\mathcal{D}}(\hat{\nu}_a(t)), \nu\right) > \frac{f(t)}{N_a(t)}$. Consequently, $\mathcal{K}_{\inf}(\Pi_{\mathcal{D}}(\hat{\nu}_a(t)), \mu) \geq \frac{f(t)}{N_a(t)}$.

We now describe two settings in which the algorithm that we have described achieves the optimal asymptotic regret bound. These two settings and the presentation thereof follows Cappé et al. [2013a]. The first family of distributions we consider for $\mathcal{D}$ is a canonical one-dimensional exponential family $\mathcal{E}$. For some dominating measure λ (not necessarily Lebesgue), open set $H \subseteq \mathbb{R}$, and twice-differentiable

[2] If $c_a \leq B$ for all a, then it is always possible to do this with K draws and respect the budget. In particular, at time $t = a$, draw arm a with probability one. If, for some a, $c_a > B$, then this strategy will violate the budget constraint, though a stochastic strategy that draws these arms a with probability c_a/B until the (random) stopping time at which the first draw occurs would respect the budget. Whether we use this strategy or just pull each arm once has essentially no effect on our analysis, and so for simplicity we assume the agent draws each arm once to initialize the algorithm.

[3] While presumably not necessary, this restriction aids our arguments in Section 6.1, and seems very mild given that at ρ the agent is indifferent to whether she pulls an arm or a pseudo-arm.

[4] An easy choice is to make $Q(t)$ a product measure with marginal probabilities $q_1(t), \ldots, q_{K+1}(t)$, but this choice is not necessary, and more careful choices may reduce the probability of overspending the budget at any given time point.

strictly convex function $b : H \to \mathbb{R}$, $\mathcal{E}$ is a set of distributions ν_η such that

$$\frac{d\nu_\eta}{d\lambda}(x) = \exp\left[x\eta - b(\eta)\right].$$

We assume that the open set H is the natural parameter space, i.e. the set of all $\eta \in \mathbb{R}$ such that $\int \exp(x\eta) d\lambda(x) < \infty$. We define the corresponding (open) set of expectations by $I \equiv \{E(\nu_\eta) : \eta \in H\} \equiv (\mu_-, \mu_+)$ and its closure by $\bar{I} = [\mu_-, \mu_+]$. We have omitted the dependence of $\mathcal{E}$ on λ and b in the notation. It is easily verified that $\mathcal{K}_{\inf}(\mu_a, c_a\rho^\star) = \mathrm{KL}(\nu_a, c_a\rho^\star)$.

For the moment suppose that $\hat{\nu}_a(t)$ is such that $\hat{\mu}_a(t) \in I$. In this case we let $\Pi_{\mathcal{D}}$ denote the maximum likelihood operator so that $\Pi_{\mathcal{D}}\left(\hat{\nu}_a(t)\right)$ returns the unique distribution in $\mathcal{D}$ indexed by the η satisfying $b'(\eta) = \hat{\mu}_a(t)$. Thus, in this setting where $\hat{\mu}_a(t) \in I$, the UCB $U_a(t)$ then takes the form of the expression in (16).

More generally, we must deal with the case that $\hat{\mu}_a(t)$ equals μ_+ or μ_-. For $\mu \in I$, define by convention $\mathrm{KL}(\mu_-, \mu) = \lim_{\mu' \to \mu_-} \mathrm{KL}(\mu_-, \mu)$, $\mathrm{KL}(\mu_+, \mu) = \lim_{\mu' \to \mu_+} \mathrm{KL}(\mu', \mu)$, and analogously for $\mathrm{KL}(\mu, \mu_-)$ and $\mathrm{KL}(\mu, \mu_+)$. Finally, define $\mathrm{KL}(\mu_-, \mu_-)$ and $\mathrm{KL}(\mu_+, \mu_+)$ to be zero. This then gives the following general expression for $U_a(t)$ that we use to replace (16) in the KL-UCB Algorithm:

$$U_a(t) \equiv \sup\left\{\mu \in \bar{I} : \mathrm{KL}\left(\hat{\mu}_a(t), \mu\right) \leq \frac{f(t)}{N_a(t)}\right\}, \; a = 1, \ldots, K. \tag{17}$$

Note that this definition of $U_a(t)$ does not explicitly include a mapping $\Pi_{\mathcal{D}}$ mapping any empirical distribution function to an element of the model $\mathcal{D}$. Thus we have avoided any problems that could arise in defining such a mapping when $\hat{\mu}_a(t)$ falls on the boundary of $\bar{I}$. The above optimization problem can be solved by noting that $\mu \mapsto \mathrm{KL}\left(\hat{\mu}_a(t), \mu\right)$ is convex, and so one can first identify the μ_0 minimizing this function, and then perform a root-finding method for monotone functions to (approximately) identify the largest $\mu \geq \mu_0$ at which $\mathrm{KL}\left(\hat{\mu}_a(t), \mu\right) - \frac{f(t)}{N_a(t)} = 0$.

The KL-UCB variant that we have presented achieves the asymptotic regret bound in the setting where $\mathcal{D} = \mathcal{E}$.

Theorem 6 (Optimality for single parameter exponential families)**.** *Suppose that $\mathcal{D} = \mathcal{E}$. Further let $f(t) = \log t + 3\log\log t$ for $t \geq 3$ and $f(1) = f(2) = f(3)$. This variant of KL-UCB satisfies (11), (12), (13) and (14). Thus, KL-UCB achieves the asymptotic regret lower bound (10) for uniformly efficient algorithms.*

Another interesting family of distributions for $\mathcal{D}$ is a set $\mathcal{B}$ of distributions on $[0, 1]$ with finite support. If the support of $\mathcal{D}$ is instead bounded in some $[-M, M]$, then the observations can be rescaled to $[0, 1]$ when selecting which arm to pull using the linear transformation $x \mapsto (x + M)/(2M)$.

If $\mathcal{D}$ is equal to $\mathcal{B}$, then Cappé et al. [2013a] observe that (16) rewrites as

$$U_a(t) = \sup\left\{E(\nu) : \mathrm{Support}[\nu] \subseteq \mathrm{Support}\left[\hat{\nu}_a(t)\right] \cup \{1\} \text{ and } \mathrm{KL}\left(\hat{\nu}_a(t), \nu\right) \leq \frac{f(t)}{N_a(t)}\right\}$$

where, for a measure ν', we use $\mathrm{Support}[\nu']$ to denote the support of ν'. They furthermore observe that this expression admits an explicit solution via the method of Lagrange multipliers.

Theorem 7 (Optimality for finitely supported distributions)**.** *Suppose that $\mathcal{D} = \mathcal{B}$. Let $\Pi_{\mathcal{D}}$ denote the identity map and $f(t) = \log t + \log\log t$ for $t \geq 2$ and $f(1) = f(2)$. Suppose that $\mu_a \in (0, 1)$ for all $a = 1, \ldots, K$. The variant of KL-UCB satisfies (11), (12), (13) and (14). Thus, KL-UCB achieves the asymptotic regret lower bound (10) for uniformly efficient algorithms.*

In both theorems, the little-oh notation hides the problem-dependent but T-independent quantities. In the proofs of Theorems 6 and 7 we refer to equations in Cappé et al. [2013b] where the reader can find explicit finite-sample, problem-dependent expressions for the $o(\log T)$ term in (11) for the settings

of Theorems 6 and 7. The argument used to establish (12) considers similar $o(\log T)$ terms to those that appear in the proof of (11), though the simplest argument for establishing (12) (which, for brevity, is the one that we have elected to present here) invokes asymptotics. The argument used to establish (13) in these settings, on the other hand, seems to be fundamentally asymptotic and does not appear to easily yield finite sample constants. Nonetheless, this is to our knowledge the first handling of thick margins in the multiple-play bandit literature, and so we believe that our rate- and constant-optimal regret guarantee is of interest despite its asymptotic nature.

Moreover, though not presented in detail here, our proof techniques can be used to establish a finite-time regret guarantee that is rate-optimal, namely is $O(\log T)$, but is constant-suboptimal. To obtain this bound, we note that, by Proposition 3, it suffices to combine (i) the previously-discussed finite-time variants of (11) and (12) that can result from the proof of Theorem 7 and (ii) the following finite-time variant of (13), which must hold for all $T \geq 1$ and some $C > 0$:

$$\text{for arms } a^\star \in \mathcal{L}: \quad \mathbb{E}_{\mathcal{V}}[N_{a^\star}(t)] = T - C \log T. \tag{18}$$

This guarantee is asymptotically weaker than that in (13) in the sense that the $o(\log T)$ term has been replaced by $O(\log T)$, but is stronger than (13) in the sense that we require a finite-time bound on the $O(\log T)$ term rather than only an asymptotic guarantee. Though we did not explicitly establish the above in our proof of Theorem 7, only a minor modification to the proof is needed. Specifically, by (29), it suffices to obtain a finite-time upper bound on $\mathbb{E}[M_a^{a^\star}(T)]$ for all $a \in \mathcal{M} \cup \mathcal{N}$ and $a^\star \in \mathcal{L}$. This upper bound can be found by noting that the proof of Lemma 18 shows that $\mathbb{E}[M_a^{a^\star}(T)] \leq O(\log T)$, and explicit finite-sample constants can be computed for this bound just as they can for (11). Plugging this into (29) then establishes (18), which in turn establishes a finite-time $O(\log T)$ regret bound. This finite-time regret bound will be valid even if $\mathcal{M}$ contains more than one arm.

4.2 Thompson Sampling

Algorithm Thompson Sampling

Parameters For each arm $a = 1, \ldots, K$, let $\Pi_a(0)$ be a prior distribution on μ_a.
for $t = 0, 1, \ldots$ **do**
  For each arm $a = 1, \ldots, K$, draw $\theta_a(t) \sim \Pi_a(t)$.
  Let $\hat{\rho}^\star(t) \equiv \rho^\star\left((\theta_a(t) : a = 1, \ldots, K)\right)$.
  For $a \in \{1, \ldots, K\}$, let $q_a(t) = \mathbb{1}\{\theta_a(t) > c_a \hat{\rho}^\star(t)\}$.
  if $\widehat{\mathcal{M}}(t) \equiv \{a : \theta_a(t) = c_a \hat{\rho}^\star(t)\}$ is non-empty **then**
    For $a \in \widehat{\mathcal{M}}(t)$, let $q_a(t) = \left[B - \sum_{a:\theta_a(t) > c_a \hat{\rho}^\star(t)} c_a\right] / \sum_{a \in \widehat{\mathcal{M}}(t)} c_a$.
  Draw $\hat{\mathcal{A}}(t+1)$ from any distribution $Q(t)$ with marginal probabilities $q_a(t)$.
  Draw the corresponding rewards $Y_a(t+1)$, $a \in \hat{\mathcal{A}}(t+1)$.
  For each $a \in \hat{\mathcal{A}}(t+1)$, obtain a new posterior $\Pi_a(t+1)$ by updating $\Pi_a(t)$ with the observation $Y_a(t+1)$.
  For each $a \notin \hat{\mathcal{A}}(t+1)$, let $\Pi_a(t+1) = \Pi_a(t)$.

Thompson sampling uses Bayesian ideas to account for the uncertainty in the estimated reward distributions. In a classical bandit setting, one first posits a (typically non-informative) prior over the means of the reward distributions, and then at each time updates the posterior and takes a random draw of the K means from the posterior and pulls the arm whose posterior draw is the largest. In our setting, this corresponds to drawing the subset of arms for which the posterior draw to cost ratio is largest (up until the budget constraint is met), which generalizes the idea initially proposed by Thompson [1933]. In the above algorithm, we focus on independent priors so that the only posteriors updated at time $(t+1)$ are those of arms in $\hat{\mathcal{A}}(t+1)$. At time $(t+1)$, Thompson Sampling first draws one sample $\theta_a(t)$ from the posterior distribution on the mean of each arm a, and then selects a subset according an oracle strategy assuming $(\theta_a(t))_{a=1,\ldots,K}$ are the true parameters.

We prove the optimality of Thompson sampling for Bernoulli rewards, for the particular choice of a uniform prior distribution on the mean of each arm. Note that the algorithm is easy to implement in that case, since $\Pi_a(t)$ is a Beta distribution with parameters $N_a(t)\hat{\mu}_a(t)+1$ and $N_a(t)(1-\hat{\mu}_a(t))+1$. Our proof relies on the same techniques as those used to prove the optimality of Thompson sampling in the standard bandit setting for Bernoulli rewards by Agrawal and Goyal [2012]. We note that Komiyama et al. [2015] also made use of some of the techniques in Agrawal and Goyal [2012] to prove the optimality of Thompson sampling for Bernoulli rewards in the multiple-play bandit setting.

Theorem 8 (Optimality for Bernoulli rewards)**.** *If the reward distributions are Bernoulli and $\Pi_a(0)$ is a standard uniform distribution for each a, then Thompson sampling satisfies (11), (12), (13) and (14). Thus, Thompson sampling achieves the asymptotic regret lower bound (10) for uniformly efficient algorithms.*

For any $\epsilon > 0$ and $a \in \underline{\mathcal{N}}$, the proof shows that Thompson sampling satisfies

$$\mathbb{E}[N_a(T)] \leq (1+\epsilon)^2 \frac{f(T)}{\mathrm{KL}(\mu_a, c_a\rho^\star)} + o(\log T).$$

The proof gives an explicit bound on the $o(\log T)$ term that depends on both the problem and the choice of ϵ. We conclude by noting that, similarly as for KL-UCB, our proof techniques can be easily adapted to give a rate-optimal but constant-suboptimal finite-time regret bound, where this bound will be valid even if $\mathcal{M}$ contains more than one arm.

5 Numerical Experiments

We now run four simulations to evaluate our theoretical results in practice, all with Bernoulli reward distributions, a horizon of $T = 100\,000$, and $K = 5$. The simulation settings are displayed in Table 1. Simulations 1-3 are run using 5 000 Monte Carlo repetitions, and Simulation 4 was run using 50 000 repetitions to reduce Monte Carlo uncertainty.

	μ	c	B	ρ	$\mathcal{L}$	$\mathcal{M}$	$\overline{\mathcal{N}}$
Sim 1	$(0.5, 0.45, 0.45, 0.4, 0.3)$	$(1, 1, 1, 1, 1)$	2	0	$\{1\}$	$\{2, 3\}$	$\emptyset$
Sim 2	$(0.7, 0.6, 0.5, 0.3, 0.2)$	$(1, 1, 1, 1, 1)$	3	0	$\{1, 2\}$	$\{3\}$	$\emptyset$
Sim 3	$(0.5, 0.45, 0.45, 0.4, 0.3)$	$(0.8, 1, 1, 0.8, 0.6)$	2	0.5	$\{1\}$	$\{4, 5, 6\}$	$\emptyset$
Sim 4	$(0.7, 0.6, 0.5, 0.3, 0.2)$	$(1.5, 1, 1, 1, 2.5)$	3	0.4	$\{2, 3\}$	$\{1\}$	$\{5\}$

Table 1: Simulation settings considered. Simulations 1 and 3 have non-unique margins so that q_a must be less than one for at least one arm $a \in \mathcal{M}$ for the budget constraint to be satisfied. In Simulation 3, the pseudo-arm $(K+1) = 6$ is in $\mathcal{M}$, and in Simulation 4 arm 5 is in $\overline{\mathcal{N}}$.

For $d \in \mathbb{R}$, we define the KL-UCB d algorithm as the instance of KL-UCB using the function $f(t) = \log t + d \log\log t$. Note that the use of both KL-UCB 3 and KL-UCB 1 are theoretically justified by the results of Theorems 6 and 7, as Bernoulli distributions satisfy the conditions of both theorems. In the settings of Simulations 1 and 2, which represent multiple-play bandit instances as B is an integer in $[1, K]$ and the cost of pulling each arm is one, we compare Thompson sampling and KL-UCB to the ESCB algorithm of Combes et al. [2015b]. As quickly explained earlier, ESCB is a generalization of the KL-UCB algorithm, designed for the combinatorial semi-bandit setting (that includes multiple-play). This algorithm computes an upper confidence bound for the sum of the arm means for each of the $\binom{K}{B}$ candidate sets $\mathcal{S}$, defined by the optimal value to

$$\sup_{(\mu_1,\dots,\mu_K)\in[0,1]^K} \sum_{a\in\mathcal{S}} \mu_K \;\text{ subject to }\; \sum_{a\in\mathcal{S}} N_a(t)\,\mathrm{KL}\left(\hat{\mu}_a(t), \mu_a\right) \leq f(t) \tag{19}$$

and draws the arms in the set $\mathcal{S}$ with the maximal index. Just like KL-UCB, ESCB uses confidence bounds whose level rely on a function f such that $f(t) \approx \log t$. Because the optimization problem solved to compute the indices (17) and (19) are different, the f functions used by KL-UCB and ESCB are not directly comparable. Nonetheless, a side-by-side comparison of the two algorithms seems to indicate that $f(t) = \log t + cB \log\log t$ for ESCB is comparable to $f(t) = \log t + c \log\log t$ for KL-UCB. Combes et al. prove an $O(\log T)$ regret bound (with a sub-optimal constant) for the version of ESCB corresponding to the constant $c = 4$, that we refer to as ESCB $4B$.

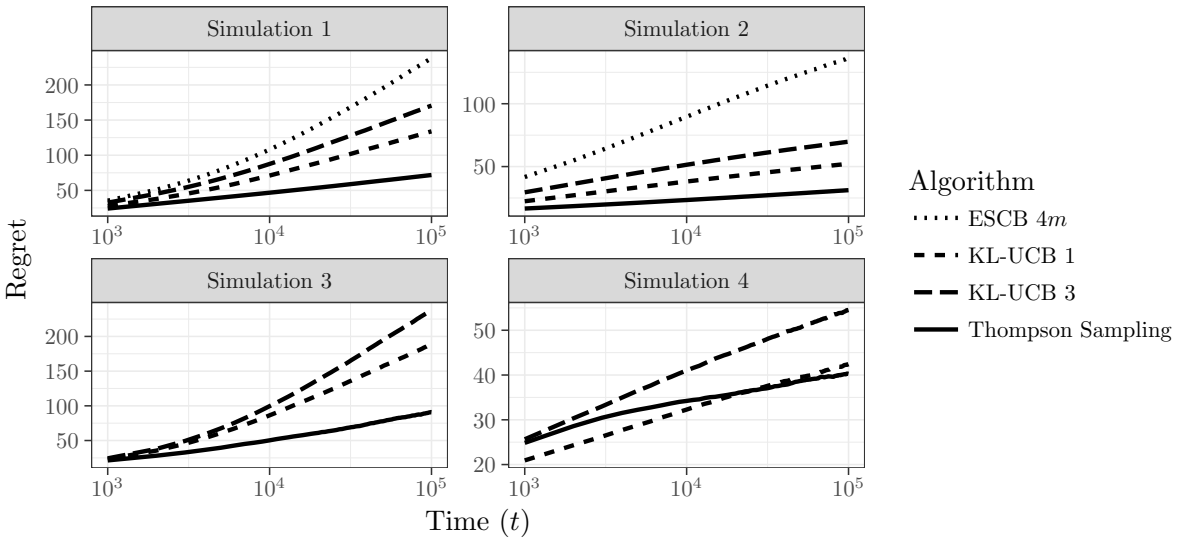

Figure 1: Regret of the four algorithms with theoretical guarantees. ESCB only run for Simulations 1 and 2 for which the cost is identically one for all arms.

Figure 1 displays the regret of the four algorithms with theoretical guarantees. All but ESCB $4B$ have been proven to be asymptotically optimal, and thus are guaranteed to achieve the theoretical lower bound asymptotically. In our finite sample simulation, Thompson sampling performs better than this theoretical guarantee may suggest (the regret lower bounds at time $T = 100\,000$ are approximately equal to 150 and 45 in Simulations 1 and 2, respectively). Indeed, Thompson sampling outperforms the KL-UCB algorithms in all but Simulation 4, while KL-UCB 1 outperforms KL-UCB 3 and KL-UCB 3 outperforms ESCB $4B$ in Simulations 1 and 2. To give the reader intuition on the relative performance of KL-UCB variants, note that in the proofs of Theorems 6 and 7 we prove that the number of pulls on each suboptimal arm a is upper bounded by $f(T)/\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star) + o(\log T)$, with an explicit finite sample constant for the $o(\log T)$ term. While $f(T) = \log T + o(\log T)$ for KL-UCB 1 and KL-UCB 3, for finite T the quantities $\log T$ and $\log T + c\log\log T$, $c = 1, 3$, are quite different. At $T = 10^5$, $\log T + \log\log T$ is 20% larger than $\log T$, and $\log T + 3\log\log T$ is 60% larger. This difference does not decay quickly with sample size: at $T = 10^{15}$, these two quantities are still respectively 10% and 30% larger than $\log T$. This makes clear the practical benefit to choosing $f(t)$ as close to $\log t$ as is theoretically justifiable: for Bernoullis, the choice of $f(t)$ in Theorem 7 yields much better results than the choice of $f(t)$ in Theorem 6.

We also compared the performance of KL-UCB 0 and ESCB 0 in Simulations 1 and 2 (details omitted here, but the exact results of this simulation are given in Figure 2 of the earlier technical report Luedtke et al., 2016). Though not theoretically justified, this choice of $f(t) = \log t$ has been used quite a lot in practice. The ordering of the three algorithms is the same in Simulations 1 and 2: Thompson Sampling performs best while ESCB 0 slightly outperforms KL-UCB 0. This should however be mitigated by the gap of numerical complexity between the two algorithms, especially when B and K are large and B/K is

not close to 0 or 1: while KL-UCB only requires running K univariate root-finding procedures regardless of B, the current proposed ESCB algorithm requires running $\binom{K}{B}$ univariate root-finding procedures. For $K = 100$ and $B = 10$, this is a difference of running 100 root-finding procedures versus more than 10^{13} of them.

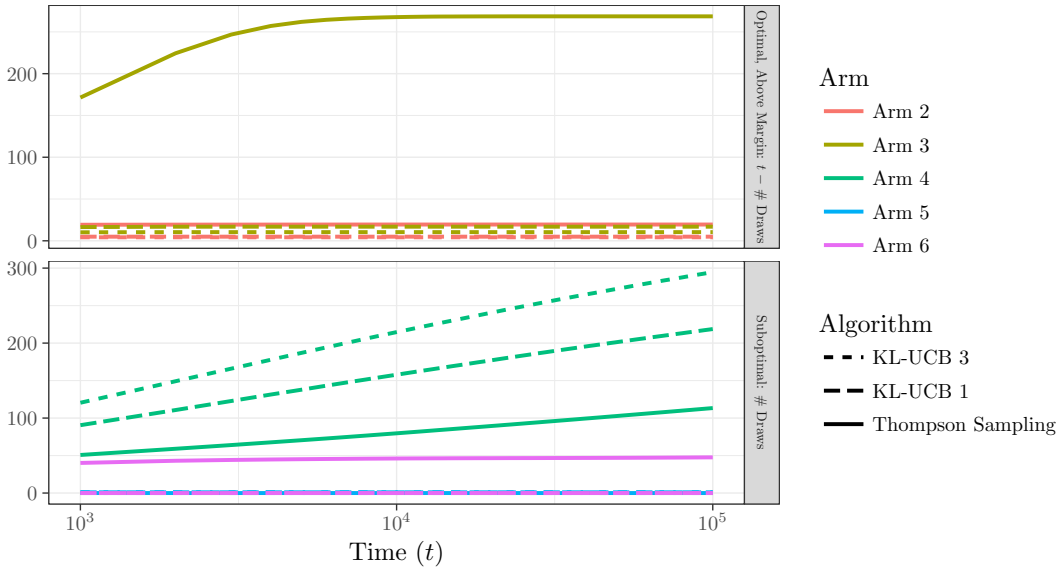

Figure 2: Time minus the number of optimal arm draws (top) and number of suboptimal arm draws (bottom) in Simulation 4.

Figure 2 displays the number of optimal and suboptimal arm draws in Simulation 4. None of the algorithms pulled the arm in $\overline{\mathcal{N}}$ (arm 5) often. Thompson Sampling pulled the indifference point pseudo-arm surprisingly often in the first 10^3 draws, and as a result arm 3 (above the margin) was also not pulled as often as would be expected in these early draws. By time 10^4, the regret of Thompson sampling appears to have stabilized, and soon outperforms that of the two KL-UCB algorithms. We also checked what would happen if the indifference point were increased from 0.4 to 0.45 (details not shown). In this case, it takes even longer for the algorithm to differentiate between arm 3 (with $\rho_3 = 0.5$) and the pseudo-arm, though by time 10^5 the algorithm again appears to have succeeded in learning that pulling arm 3 is to be preferred over pulling the peudo-arm.

6 Proofs of Optimality of KL-UCB and Thompson Sampling

We now outline our proofs of optimality for the KL-UCB and Thompson sampling schemes. We break this section into three subsections. Section 6.1 establishes that the arms in $\mathcal{N}$, i.e. the suboptimal arms, are not pulled often (satisfy Equations 11 and 12). Due to the differences in proof methods, we consider the KL-UCB and Thompson sampling schemes separately in this subsection. Section 6.2 justifies that when $\rho^\star > \rho$, the budget constraint is most often saturated, that is the third term in the regret is negligible. Finally Section 6.3 establishes that the arms in $\mathcal{L}$, i.e. the optimal arms away from the margin, are pulled often (satisfy Equation 13). We give the outline of the proofs for the KL-UCB and Thompson sampling schemes simultaneously, though we provide the detailed arguments separately in Appendices C and D, respectively. We note that the order of presentation of the two subsections is important: the arguments used in Section 6.3 rely on the validity of (11) and (12), which is established in Section 6.1.

To ease the presentation, we find it convenient to consider the extended bandit model presented in Section 2.2, in which a pseudo-arm $K + 1$ of cost B is added to the bandit instance, with a positive

probability of pulling arm $K+1$ representing the decision not to spend the entire budget on pulling arms $1,\ldots,K$. Though both the KL-UCB and Thompson Sampling algorithms were presented without this extra arm, we already noted that for each t, $q_{K+1}(t) = 1 - \frac{1}{B}\sum_{a=1}^{K} c_a q_a(t)$. The UCB index $U_{K+1}(t)$ and posterior draw $\theta_{K+1}(t)$ for arm $K+1$ are both equal to $B\rho$ for all t. For the sake of condensing notation in our study of (expected) regret, it will be convenient to consider a hypothetical scenario in which arm $K+1$ is pulled with probability $q_{K+1}(t)$ at each time point, even though the outcome of these pulls has no effect on the behavior of the algorithms.

6.1 Suboptimal arms not pulled often

In this section, we establish (11) and (12) for KL-UCB and Thompson Sampling.

For a fixed arm a, the KL-UCB and Thompson sampling proofs will both rely on a quantity $\rho^\dagger \in (\rho_a, \mu_+/c_a)$, though we note that the value that we select for $\rho^\dagger$ will vary between the proofs.

KL-UCB

Preliminary: a general analysis. We start by giving a general analysis of KL-UCB in our setting, and then use it to prove Theorems 6 and 7. Fix $a \in \mathcal{N}\backslash\{K+1\}$. The arguments in this section generalize those given in Cappé et al. [2013a,b] for the case where one arm is drawn at each time point and there is no budget constraint. Let $\mu^\dagger \in (\mu_a, \mu_+)$ be some real number. If $a \in \underline{\mathcal{N}}$, then we will choose $\mu^\dagger = c_a\rho^\star$. If, on the other hand, $a \in \overline{\mathcal{N}}$, then we will choose $\mu^\dagger$ to be less than μ_+. Let $\rho^\dagger$ be a constant that is either equal to or slightly less than $\mu^\dagger/c_a$. Below we take minimums over $a^\star \in \mathcal{S} \equiv (\mathcal{L}\cup\mathcal{M})\backslash\{K+1\}$: if $\mathcal{S} = \emptyset$, then we take these minimums to be equal to negative infinity. When we later take sums over $a^\star \in \mathcal{S}$, we let empty sums equal zero.

We now establish that, for all $t \geq K$,

$$\left\{a \in \hat{\mathcal{A}}(t+1)\right\} \subseteq \left[\cup_{a^\star\in\mathcal{S}}\left\{c_{a^\star}\rho^\dagger \geq U_{a^\star}(t)\right\}\right] \cup \left\{a \in \hat{\mathcal{A}}(t+1), c_a\rho^\dagger < U_a(t)\right\}. \tag{20}$$

We separately handle the cases that $\rho^\star > \rho$ and $\rho^\star = \rho$. If $\rho^\star > \rho$, playing all of the arms in $\mathcal{S}$ would spend at least the allotted budget B. Hence, on the event $\left\{\forall a^\star \in \mathcal{S}, U_{a^\star}(t)/c_{a^\star} > \rho^\dagger\right\}$, it holds that $\hat{\rho}^\star(t) > \rho^\dagger$. If moreover $a \in \hat{\mathcal{A}}(t+1)$, one has $U_a(t) \geq c_a\hat{\rho}^\star(t) > c_a\rho^\dagger$. If $\rho = \rho^\star$, it holds that $\{a \in \hat{\mathcal{A}}(t+1)\} \subseteq \{a \in \hat{\mathcal{A}}(t+1), c_a\rho^\dagger < U_a(t)\}$. Indeed, if $\hat{\rho}^\star(t) > \rho$ the algorithm only pulls arms a if $U_a(t) \geq \hat{\rho}^\star(t)c_a > \rho c_a$ and if $\hat{\rho}^\star(t) = \rho$, then the algorithm only pulls arm a if $U_a(t) > c_a\rho$, see Footnote [3]. As $\rho^\dagger$ is smaller or equal to $\rho^\star = \rho$, it follows that $U_a(t) > c_a\rho^\dagger$ in both cases.

For each $\zeta > 0$ and $\tilde{\mu} < \mu_+$, we now introduce the set $\mathcal{C}_{\tilde{\mu},\zeta}$. In the setting of Theorem 6,

$$\mathcal{C}_{\tilde{\mu},\zeta} \equiv \left\{\nu' : \text{Support}[\nu'] \subseteq \bar{I}\right\} \cap \left\{\nu' : \exists\, \mu \in (\tilde{\mu}, \mu_+] \text{ with } \mathrm{KL}(E(\nu'), \mu) \leq \zeta\right\},$$

where above $\mathrm{KL}(E(\nu'), \mu)$ is the KL-divergence in the canonical exponential family $\mathcal{E}$. In the setting of Theorem 7,

$$\mathcal{C}_{\tilde{\mu},\zeta} \equiv \{\nu' : \text{Support}[\nu'] \subseteq [0,1]\} \cap \{\nu' : \exists\, \nu \in \mathcal{B} \text{ with } \tilde{\mu} < E(\nu) \text{ and } \mathrm{KL}(\Pi_{\mathcal{D}}(\nu'), \nu) \leq \zeta\}.$$

In both settings, we will invoke this set at $\tilde{\mu} = c_a\rho^\dagger < \mu_+$. The set $\mathcal{C}_{\tilde{\mu},\zeta}$ is defined in both settings so that $\tilde{\mu} < U_a(t)$ if and only if $\hat{\nu}_a(t) \in \mathcal{C}_{\tilde{\mu},f(t)/N_a(t)}$. Recalling that $\mathbb{E}[N_a(T)] = \sum_{t=0}^{T-1}\mathbb{P}\{a \in \mathcal{A}(t+1)\}$, a union bound gives

$$\mathbb{E}[N_a(T)] \leq 1 + \sum_{a^\star\in\mathcal{S}}\sum_{t=K}^{T-1}\mathbb{P}\left\{c_{a^\star}\rho^\dagger \geq U_{a^\star}(t)\right\} + \sum_{t=K}^{T-1}\mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{c_a\rho^\dagger, f(t)/N_a(t)}\right\}.$$

In analogue to Equation 8 in Cappé et al. [2013a], the above rightmost term satisfies

$$\sum_{t=K}^{T-1}\mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{c_a\rho^\dagger, f(t)/N_a(t)}\right\}$$

$$
\begin{aligned}
&\leq \sum_{t=K}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{c_a\rho^\dagger, f(T)/N_a(t)}\right\} \\
&= \sum_{t=K}^{T-1} \sum_{n=2}^{T-K+1} \mathbb{P}\left\{\hat{\nu}_{a,n-1} \in \mathcal{C}_{c_a\rho^\dagger, f(T)/(n-1)}, \tau_{a,n} = t+1\right\} \qquad (21) \\
&\leq \sum_{n=1}^{T-K} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{c_a\rho^\dagger, f(T)/n}\right\},
\end{aligned}
$$

where the final inequality holds because, for each n, $\tau_{a,n} = t+1$ for at most one t in $\{K, \ldots, T-1\}$. We will upper bound the terms with $n = 1, \ldots, b_a^\star(T)$ in the sum on the right by 1, where

$$
b_a^\star(T) \equiv \left\lceil \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger)} \right\rceil \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger)} + 1.
$$

This gives the bound

$$
\sum_{n=1}^{T-K} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{c_a\rho^\dagger, f(T)/n}\right\} \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger)} + 1 + \sum_{n=b_a^\star(T)+1}^{\infty} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{c_a\rho^\dagger, f(T)/n}\right\}.
$$

Hence,

$$
\mathbb{E}[N_a(T)] \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger)} + \underbrace{\sum_{n=b_a^\star(T)+1}^{\infty} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{c_a\rho^\dagger, f(T)/n}\right\}}_{\text{Term 1}} + \sum_{a^\star \in \mathcal{S}} \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\left\{c_{a^\star}\rho^\dagger \geq U_{a^\star}(t)\right\}}_{\text{Term } 2a^\star} + 2. \qquad (22)
$$

Up until this point we have not committed to any particular choice of $\mu^\dagger$, $\rho^\dagger$, or non-decreasing function $f : \mathbb{N} \to \mathbb{R}$. We now give proofs of (11) and (12) in the settings of Theorems 6 and 7. For each proof we use the choice of f from the theorem statement and make particular choices of $\mu^\dagger$ and $\rho^\dagger$.

Lemma 9. *Eq. 11 holds in the settings of Theorems 6 and 7*

Proof. Fix $a \in \mathcal{N} \backslash \{K+1\}$. If $a \in \underline{\mathcal{N}}$, then let $\mu^\dagger = c_a\rho^\star$ and, if $a \in \overline{\mathcal{N}}$, then let $\mu^\dagger \in (\mu_a, \mu_+)$. In the setting of Theorem 6 let $\rho^\dagger = \mu^\dagger/c_a$ and in the setting of Theorem 7 let $\rho^\dagger = \left[1 - \log(T)^{-1/5}\right]\mu^\dagger/c_a$. Lemma A.1 shows that Term 1 is $o(\log T)$ and includes references on where to find an explicit finite sample upper bound, where this upper bound will rely on the choice of $\mu^\dagger < \mu_+$ if $a \in \overline{\mathcal{N}}$. Fix $a^\star \in \mathcal{S}$. Noting that $\rho^\dagger \leq \left[1 - \log(T)^{-1/5}\right]\rho_{a^\star}$ (Theorem 6) and $\rho^\dagger \leq \rho_{a^\star}$ (Theorem 7), Term $2a^\star$ is $o(\log T)$ in both settings by Lemma A.2, with an exact finite sample upper bound given in the proof thereof. Thus, $\sum_{a^\star \in \mathcal{S}} \text{Term } 2a^\star = o(\log T)$. This completes the proof of (11). □ □

Lemma 10. *Eq. 12 holds in the settings of Theorems 6 and 7*

Proof. For $a \in \overline{\mathcal{N}}$, so far we have established that, for arbitrary $\mu^\dagger \in (\mu_a, \mu_+)$,

$$
\mathbb{E}[N_a(T)] \leq \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger)} + r(T, \mu^\dagger),
$$

where $r(T, \mu^\dagger)/\log T \to 0$ for fixed $\mu^\dagger$. As this holds for every $\mu^\dagger$, there exists a sequence $\mu^\dagger(T) \to \mu_+$ such that $r(T, \mu^\dagger(T))/\log T \to 0$. In both settings $\liminf_{\mu^\dagger \to \mu_+} \mathcal{K}_{\inf}(\nu_a, \mu^\dagger) = +\infty$, and so using this $\mu^\dagger(T)$ sequence shows that $\mathbb{E}[N_a(T)] = o(\log T)$. □ □

Thompson Sampling

This proof is inspired by the analysis of Thompson sampling proposed by Agrawal and Goyal [2012]. We work with a suboptimal arm $a \in \mathcal{N}\backslash\{K+1\}$ in most of this section, though we state one of the results (Lemma 11) for general arms $a \in \{1, \ldots, K+1\}$ since it will prove useful later. We will let $\rho^\dagger$ and $\rho^\ddagger$ be numbers (to be specified later) satisfying $\rho_a < \rho^\dagger < \rho^\ddagger < 1/c_a$. Observe that $\left\{a \in \hat{\mathcal{A}}(t+1)\right\}$ equals

$$\begin{aligned}
&\left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) \le c_a\rho^\ddagger\right\} \cup \left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) > c_a\rho^\ddagger\right\} \\
&\subseteq \left[\cup_{a^\star \in \mathcal{L}\cup\mathcal{M}} \left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) \le c_a\rho^\ddagger, \theta_{a^\star}(t) \le c_{a^\star}\hat{\rho}^\star\right\}\right] \cup \left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) > c_a\rho^\ddagger\right\}.
\end{aligned}$$

By the absolute continuity of the beta distribution, with probability one at most one $a' \in \{1, \ldots, K+1\}$ satisfies $\theta_{a'}(t) = c_{a'}\hat{\rho}^\star$, and hence, conditional on $\mathcal{F}(t)$, the leading event above is almost surely equivalent to the event

$$\cup_{a^\star \in \mathcal{L}\cup\mathcal{M}} \left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) \le c_a\rho^\ddagger, \theta_{a^\star}(t) < c_{a^\star}\hat{\rho}^\star\right\}.$$

If $K+1 \in \mathcal{M}$, then the fact that $a \in \hat{\mathcal{A}}(t+1)$ implies that $\theta_a(t)/c_a(t) \ge \hat{\rho}^\star(t)$ shows that the event in the union above at $a^\star = K+1$ never occurs, since on this event $\rho_{K+1} = \theta_{K+1}(t)/c_{K+1} < \rho^\ddagger$, which contradicts our choice that $\rho^\ddagger < \rho^\star = \rho_{K+1}$. Hence, the union above can be taken over $\mathcal{S}$ regardless of whether or not $K+1 \in \mathcal{M}$. Furthermore,

$$\begin{aligned}
&\left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) > c_a\rho^\ddagger\right\} \\
&\subseteq \left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) > c_a\rho^\ddagger, \hat{\mu}_a(t) \le c_a\rho^\dagger\right\} \cup \left\{a \in \hat{\mathcal{A}}(t+1), \hat{\mu}_a(t) > c_a\rho^\dagger\right\}.
\end{aligned}$$

Recalling that $\mathbb{E}[N_a(T)] = \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1)\right\}$,

$$\begin{aligned}
\mathbb{E}[N_a(T)] \le &\underbrace{\sum_{a^\star \in \mathcal{S}} \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) \le c_a\rho^\ddagger, \theta_{a^\star}(t) < c_{a^\star}\hat{\rho}^\star\right\}}_{\text{Term I}a^\star} \\
&+ \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) > c_a\rho^\ddagger, \hat{\mu}_a(t) \le c_a\rho^\dagger\right\}}_{\text{Term II}} \\
&+ \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \hat{\mu}_a(t) > c_a\rho^\dagger\right\}}_{\text{Term III}}. \qquad (23)
\end{aligned}$$

The above decomposition does not depend on the algorithm. Bounding Terms I$a^\star$, $a^\star \in \mathcal{S}$, and Term II will rely on arguments that are specific to Thompson Sampling. Fix $a^\star \in \mathcal{S}$ and let $p_{a^\star}^{\rho^\ddagger}(t) \equiv \mathbb{P}(\theta_{a^\star}(t) > c_{a^\star}\rho^\ddagger \,|\, \mathcal{F}(t))$. Note that $p_{a^\star}^{\rho^\ddagger}(t) \ne p_{a^\star}^{\rho^\ddagger}(t+1)$ implies $a^\star \in \hat{\mathcal{A}}(t+1)$. Thus $p_{a^\star}^{\rho^\ddagger}(t)$ is equal to $p_{a^\star,n}^{\rho^\ddagger} \equiv p_{a^\star}^{\rho^\ddagger}(\tau_{a^\star,n})$ for all t such that $N_{a^\star}(t) = n$. We now state Lemma 11, that generalizes Lemma 1 in Agrawal and Goyal [2012].

Lemma 11. *If $a \in \{1, \ldots, K+1\}$, $a^\star \in \mathcal{S}$, and $\rho^\ddagger$ satisfies $c_{a^\star}\rho^\ddagger < 1$, then, for all $t \ge 0$,*

$$\mathbb{P}\left(a \in \hat{\mathcal{A}}(t+1), \theta_a(t) \le c_a\rho^\ddagger, \theta_{a^\star}(t) < c_{a^\star}\hat{\rho}^\star \,\middle|\, \mathcal{F}(t)\right) \le \frac{1 - p_{a^\star}^{\rho^\ddagger}(t)}{p_{a^\star}^{\rho^\ddagger}(t)} \mathbb{P}\left(\theta_{a^\star}(t)/c_{a^\star} \ge \hat{\rho}^\star(t) | \, \mathcal{F}(t)\right).$$

The proof can be found in Appendix D. Observe that the upper bound in the above lemma does not rely on a. We have another lemma, that relies on a lower bound on the probability $\mathring{q}_{a^\star}$, to be defined shortly, that is possible for $q_{a^\star}(t)$ given that $\theta_{a^\star}(t)/c_{a^\star} \geq \hat{\rho}^\star(t)$. By the absolute continuity of the beta distribution, we also have that

$$\begin{aligned}\mathbb{P}\left(a^\star \in \hat{\mathcal{A}}(t+1)\middle|\mathcal{F}(t)\right) &= \mathbb{P}\left(a^\star \in \hat{\mathcal{A}}(t+1), \frac{\theta_{a^\star}(t)}{c_{a^\star}} \geq \hat{\rho}^\star(t)\middle|\mathcal{F}(t)\right) \\ &= \mathbb{P}\left(a^\star \in \hat{\mathcal{A}}(t+1)\middle|\frac{\theta_{a^\star}(t)}{c_{a^\star}} \geq \hat{\rho}^\star(t), \mathcal{F}(t)\right)\mathbb{P}\left(\frac{\theta_{a^\star}(t)}{c_{a^\star}} \geq \hat{\rho}^\star(t)\middle|\mathcal{F}(t)\right).\end{aligned}$$

We lower bound the leading term in the product on the right by

$$\mathring{q}_{a^\star} \equiv \min\left\{1, \min_{\mathcal{H}\subseteq\{1,\ldots,K\}\setminus\{a^\star\}:\sum_{\tilde{a}\in\mathcal{H}} c_{\tilde{a}} < B} \frac{B - \sum_{\tilde{a}\in\mathcal{H}} c_{\tilde{a}}}{c_{a^\star}}\right\}.$$

Because $c_{K+1} = B$, one could equivalently take the minimum over $\mathcal{H} \subseteq \{1,\ldots,K+1\}\setminus\{a^\star\}$. To see that this is a lower bound, consider two cases. If $\theta_{a^\star}(t)/c_{a^\star} > \hat{\rho}^\star(t)$, then $a \in \hat{\mathcal{A}}(t+1)$ with probability one, and so the above is a lower bound. If $\theta_{a^\star}(t)/c_{a^\star} = \hat{\rho}^\star(t)$, then the numerator $B - \sum_{\tilde{a}\in\mathcal{H}} c_{\tilde{a}}$ of the inner minimum (over $\mathcal{H}$) above represents the minimum possible amount of remaining budget when arm $a^\star$ is the unique arm on the estimated margin. The estimated margin is almost surely (over the draws of $\theta(t)$) singleton. Clearly, $\mathring{q}_{a^\star} > 0$. As a consequence,,

$$\mathbb{P}\left(\frac{\theta_{a^\star}(t)}{c_{a^\star}} \geq \hat{\rho}^\star(t)\middle|\mathcal{F}(t)\right) \leq \mathring{q}_{a^\star}^{-1}\,\mathbb{P}\left(a^\star \in \hat{\mathcal{A}}(t+1)\middle|\mathcal{F}(t)\right). \tag{24}$$

We have the following lemma, whose proof can be found in Appendix D.

Lemma 12. *If $a^\star \in \mathcal{S}$ and $c_{a^\star}\rho^\ddagger < 1$, then, for all $t \geq 0$,*

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1 - p_{a^\star}^{\rho^\ddagger}(t)}{p_{a^\star}^{\rho^\ddagger}(t)}\,\mathbb{P}\left(\theta_{a^\star}(t)/c_{a^\star} \geq \hat{\rho}^\star(t)\middle|\,\mathcal{F}(t)\right)\right] \leq \mathring{q}_{a^\star}^{-1}\,\mathbb{E}\left[\sum_{n=0}^{T-1} \frac{1 - p_{a^\star,n}^{\rho^\ddagger}}{p_{a^\star,n}^{\rho^\ddagger}}\right].$$

Combining the two preceding lemmas yield the inequality

$$\text{Term Ia}^\star \leq \mathring{q}_{a^\star}^{-1}\,\mathbb{E}\left[\sum_{n=0}^{T-1} \frac{1 - p_{a^\star,n}^{\rho^\ddagger}}{p_{a^\star,n}^{\rho^\ddagger}}\right]. \tag{25}$$

Note crucially that we have upper bounded the sum over time on the left-hand side by a sum over the number of pulls of arm $a^\star$ on the right-hand side. There appears to be a steep price to pay for this transfer from a sum over time to a sum over counts: the right-hand side inverse weights by a conditional probability, which may be small for certain realizations of the data. Lemma 2 in Agrawal and Goyal [2012], that we restate below using our modified notation, establishes that this inverse weighting does not cause a problem for Thompson sampling with Bernoulli rewards and independent beta priors. If $\rho^\ddagger < \rho^\star$, then the proceeding lemma implies that, for each $a^\star \in \mathcal{S}$, Term Ia* is $O(1)$, i.e. is $o(\log T)$ with much to spare. Obviously, this implies that $\sum_{a^\star\in\mathcal{S}} \text{Term Ia}^\star = o(\log T)$ as well.

Lemma 13 (Lemma 2 from Agrawal and Goyal, 2012)**.** *If $a^\star \in \mathcal{S}$ and $\rho^\ddagger < \rho_{a^\star}$, then, with $\Delta \equiv \mu_{a^\star} - c_{a^\star}\rho^\ddagger$,*

$$\mathbb{E}\left[\frac{1 - p_{a^\star,n}^{\rho^\ddagger}}{p_{a^\star,n}^{\rho^\ddagger}}\right] = \begin{cases} \frac{3}{\Delta}, & \text{for } n < \frac{8}{\Delta} \\ \Theta\left(e^{-\Delta^2 n/2} + \frac{1}{(n+1)\Delta^2}e^{-\mathrm{KL}(c_{a^\star}\rho^\ddagger,\mu_{a^\star})n} + \frac{1}{\exp(\Delta^2 n/4)-1}\right), & \text{for } n \geq \frac{8}{\Delta}. \end{cases}$$

Above $\Theta(\cdot)$ is used to represent big-Theta notation.

We now turn to Term II. The following result mimics Lemma 4 in Agrawal and Goyal [2012], and is a consequence of the close link between beta and binomial distributions and the Chernoff-Hoeffding bound. We provide a proof of this result in Appendix D.

Lemma 14. *If $a \in (\mathcal{M} \cup \mathcal{N}) \backslash \{K+1\}$ and $\rho_a < \rho^\dagger < \rho^\ddagger$, where $c_a \rho^\ddagger < 1$, then*

$$\text{Term II} \equiv \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \theta_a(t) > c_a\rho^\ddagger, \hat{\mu}_a(t) \leq c_a\rho^\dagger\right\} \leq \frac{\log T}{\mathrm{KL}(c_a\rho^\dagger, c_a\rho^\ddagger)}.$$

We now turn to Term III. Note that

$$\begin{aligned}
\text{Term III} &= \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}\left\{a \in \hat{\mathcal{A}}(t+1), \hat{\mu}_{a,N_a(t)} > c_a\rho^\dagger\right\}\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{n=0}^{T-1} \mathbb{1}\left\{\tau_{a,n+1} = t+1, \hat{\mu}_{a,n} > c_a\rho^\dagger\right\}\right] \\
&\leq \sum_{n=0}^{T-1} \mathbb{P}\left\{\hat{\mu}_{a,n} > c_a\rho^\dagger\right\}, \qquad (26)
\end{aligned}$$

where the latter inequality holds because $\tau_{a,n+1} = t+1$ for at most one t in $\{0, \ldots, T-1\}$. The following lemma controls the right-hand side of the above.

Lemma 15. *Fix an arm $a \in \{1, \ldots, K\}$. If $\rho^\dagger > \rho_a$ and $c_a\rho^\dagger < 1$, then*

$$\sum_{n=0}^{T-1} \mathbb{P}\left\{\hat{\mu}_{a,n} > c_a\rho^\dagger\right\} \leq 1 + \frac{1}{\mathrm{KL}(c_a\rho^\dagger, \mu_a)}.$$

The proof is omitted, but is an immediate consequence of the Chernoff-Hoeffding bound and the additional bounding from the proof of Lemma 3 in Agrawal and Goyal [2012]. Thus we have shown that Term III is $o(\log T)$, with much to spare as well.

The proof of (11) and (12) in the setting of Theorem 8 is now straightforward.

Lemma 16. *Eq. 11 holds in the setting of Theorem 8.*

Proof. Fix $a \in \mathcal{N} \backslash \{K+1\}$. Let $\mu^\dagger = c_a\rho^\star$ if $a \in \underline{\mathcal{N}}$, and let $\mu^\dagger$ be slightly less than μ_+ if $a \in \overline{\mathcal{N}}$. Fix $\rho^\dagger < \rho^\ddagger$ and $\rho^\ddagger$ (to be specified shortly) so that $\rho_a < \rho^\dagger < \rho^\ddagger < \mu^\dagger/c_a$ and $\epsilon \in (0, 1]$ a constant. Plugging our results on each Term I$a^\star$ and on Terms II and III into (23) then yields that

$$\mathbb{E}\left[N_a(T)\right] \leq \frac{\log T}{\mathrm{KL}(c_a\rho^\dagger, c_a\rho^\ddagger)} + 1 + \frac{1}{\mathrm{KL}(c_a\rho^\dagger, \mu_a)} + O(1).$$

Select $\rho^\dagger$ so that $\mathrm{KL}(c_a\rho^\dagger, \mu^\dagger) = \frac{\mathrm{KL}(\mu_a, \mu^\dagger)}{1+\epsilon}$ and $\rho^\ddagger$ so that $\mathrm{KL}(c_a\rho^\dagger, c_a\rho^\ddagger) = \frac{\mathrm{KL}(c_a\rho^\dagger, \mu^\dagger)}{1+\epsilon}$, since this gives $\mathrm{KL}(c_a\rho^\dagger, c_a\rho^\ddagger) = \frac{\mathrm{KL}(\mu_a, \mu^\dagger)}{(1+\epsilon)^2}$. Hence,

$$\mathbb{E}[N_a(T)] \leq (1+\epsilon)^2 \frac{f(T)}{\mathrm{KL}(\mu_a, \mu^\dagger)} + r(T, \mu^\dagger),$$

where $r(T, \mu^\dagger)/\log T \to 0$ for fixed $\mu^\dagger$. □ □

Lemma 17. *Eq. 12 holds in the setting of Theorem 8.*

Proof. If $a \in \underline{\mathcal{N}}$, then dividing both sides by $\log T$, and then taking $T \to \infty$ followed by $\epsilon \to 0$ gives (11). If, on the other hand, $a \in \overline{\mathcal{N}}$, then we use that there exists a sequence $\mu^\dagger(T)$ such that $r(T, \mu^\dagger(T))/\log T \to 0$. Because $\liminf_{\mu^\dagger \to \mu_+} = +\infty$, then dividing both sides by $\log T$, taking the limit as $T \to \infty$, followed by $\epsilon \to 0$, gives (12) in the case where $a \in \overline{\mathcal{N}}$. □ □

6.2 Budget saturation when $\rho^\star > \rho$

Assuming $\rho^\star > \rho$, we prove (14) for KL-UCB and Thompson Sampling in the setting of Theorems 6 and 7 and Theorem 8 respectively. Recall that the third term in the regret decomposition (6) can be expressed in terms of the number of draws of the supplementary arm $K+1$ in the extended bandit model:

$$BT - \sum_{a=1}^{K} c_a \mathbb{E}_\nu[N_a(T)] = B\mathbb{E}[N_{K+1}(T)].$$

We prove below for each algorithm that $\mathbb{E}[N_{K+1}(T)] = o(\log(T))$, as a by product from specific elements already established when controlling the number of suboptimal draws.

KL-UCB

For any $\rho^\dagger \in (\rho, \rho^\star]$ and any $t \geq K$, it holds that, for T large enough,

$$\left\{K+1 \in \hat{\mathcal{A}}(t+1)\right\} \subseteq \bigcup_{a^\star \in \mathcal{S}} \left\{c_{a^\star}\rho \geq U_{a^\star}(t)\right\} \subseteq \bigcup_{a^\star \in \mathcal{S}} \left\{c_{a^\star}\rho^\dagger \geq U_{a^\star}(t)\right\}.$$

The first inclusion must hold because if all the arms in $\mathcal{S}$ had satisfied $U_{a^\star}/c_{a^\star} \geq \rho$, then including all of those arms in $\hat{\mathcal{A}}(t+1)$ would have been enough to saturate the budget and $K+1$ would not have been selected. The second inclusion holds because $\rho^\dagger > \rho$. Hence, $\mathbb{E}[N_{K+1}(T)] \leq \sum_{a^\star \in \mathcal{S}} \text{Term } 2a^\star$ (see Equation 22 for its definition). This condition is always satisfied by the choice $\rho^\dagger = \rho^\star$ that we have used in the setting of Theorem 6, and it holds for all T sufficiently large for the choice $\rho^\dagger = \left[1 - \log(T)^{-1/5}\right]\rho^\star$ that we have used in the setting of Theorem 7. Lemma A.2 shows that each Term $2a^\star$ is again $o(\log T)$.

Thompson Sampling

We have that

$$\{K+1 \in \hat{\mathcal{A}}(t+1)\} \subseteq \bigcup_{a^\star \in \mathcal{S}} \{K+1 \in \hat{\mathcal{A}}(t+1), \theta_{a^\star}(t) \leq c_{a^\star}\hat{\rho}^\star\}.$$

As $\rho < \rho^\star$ and $\theta_{K+1}(t) = c_{K+1}\rho$ with probability one, $\mathbb{E}[N_a(T)] \leq \sum_{a^\star \in \mathcal{S}} \text{Term Ia}^\star$ provided $\rho^\dagger \in (\rho, \rho^\star)$ (see Equation 23 for its definition). Thus, we can invoke Lemma 11 (that holds for $a = K+1$), followed by Lemmas 12 and 13, to show that $\mathbb{E}[N_{K+1}(T)] = O(1)$, and therefore is $o(\log T)$ with much to spare.

6.3 Optimal arms away from margin pulled $T - o(\log T)$ times

We now show that the optimal arms away from the margin ($a^\star \in \mathcal{L}$) are pulled often. We start by giving an analysis that applies to any algorithm that, to decide which arms to draw at time $t+1$, based on $\mathcal{F}(t)$ and possibly some external stochastic mechanism, defines indices $I_a(t)$, $a = 1, \ldots, K+1$, and then defines the threshold $\hat{\rho}^\star(t) \equiv \rho^\star(c_a I_a(t) : a = 1, \ldots, K+1)$, and, for all arms a with $I_a(t) \neq \hat{\rho}^\star(t)$, assigns mass $q_a(t) = \mathbb{1}\{I_a(t) > \hat{\rho}^\star(t)\}$. The arms with $I_a(t) = \hat{\rho}^\star(t)$ are assumed to be drawn so that $\sum_{a=1}^{K+1} c_a q_a(t) = B$. We then specialize the discussion to KL-UCB and Thompson sampling, where $I_a(t)$ is respectively equal to $U_a(t)/c_a$ and $\theta_a(t)/c_a$. For the remainder of this section, we fix an optimal arm $a^\star \in \mathcal{L}$. Observe that, for $t \geq K$ (KL-UCB) or $t \geq 0$ (Thompson sampling),

$$\begin{aligned}
\{I_{a^\star}(t) < \hat{\rho}^\star(t)\} &= \cup_{a \in \mathcal{M}\cup\mathcal{N}} \{I_a(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t)\} \\
&= \left[\cup_{a \in (\mathcal{M}\cup\mathcal{N})\setminus\{K+1\}} \{I_a(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t), I_{K+1}(t) < \hat{\rho}^\star(t)\}\right] \\
&\quad \cup \{I_{K+1}(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t)\}.
\end{aligned}$$

Recalling (24), we see that, for Thompson sampling,

$$T - \mathbb{E}[N_{a^\star}(T)] = T - \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{P}\left\{a^\star \in \hat{\mathcal{A}}(t+1) \middle| \mathcal{F}(t)\right\}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{P}\left\{a^\star \notin \hat{\mathcal{A}}(t+1) \middle| \mathcal{F}(t)\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{P}\left\{I_{a^\star}(t) < \hat{\rho}^\star(t) | \mathcal{F}(t)\right\}\right] \\
&\leq \sum_{a \in (\mathcal{M}\cup\mathcal{N})\backslash\{K+1\}} \sum_{t=0}^{T-1} \mathbb{P}\left\{I_a(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t), I_{K+1}(t) < \hat{\rho}^\star(t)\right\} \\
&\quad + \sum_{t=0}^{T-1} \mathbb{P}\left\{I_{K+1}(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t)\right\}, \qquad (27)
\end{aligned}
$$

where the first inequality holds because $\{a \notin \hat{\mathcal{A}}(t+1)\} \subseteq \{I_{a^\star}(t) < \hat{\rho}^\star(t)\}$ and the second inequality holds by the preceding display. We have a similar identity for KL-UCB, though the identity is slightly different due to the initiation of each of the K arms. Specifically,

$$
\begin{aligned}
T - K + 1 - \mathbb{E}[N_{a^\star}(T)] \leq &\sum_{a \in (\mathcal{M}\cup\mathcal{N})\backslash\{K+1\}} \sum_{t=0}^{T-1} \mathbb{P}\left\{I_a(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t), I_{K+1}(t) < \hat{\rho}^\star(t)\right\} \\
&+ \sum_{t=0}^{T-1} \mathbb{P}\left\{I_{K+1}(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t)\right\}. \qquad (28)
\end{aligned}
$$

For $a \in \mathcal{M}\cup\mathcal{N}$, let $\mathcal{H}$ denote the collection of all subsets $\mathcal{H}$ of $\{1,\ldots,K\}\backslash\{a, a^\star\}$ for which $\sum_{\tilde{a}\in\mathcal{H}} c_{\tilde{a}} < B$. For $a \in \mathcal{M}\cup\mathcal{N}$, we then define

$$
\breve{q}_a^{a^\star} \equiv \begin{cases} \min\left\{1, \min_{\mathcal{H}\in\mathcal{H}} \frac{B - \sum_{\tilde{a}\in\mathcal{H}} c_{\tilde{a}}}{c_a}\right\}, & \text{if } a = K+1 \text{ or Thompson Sampling,} \\ \min\left\{1, \min_{\mathcal{H}\in\mathcal{H}} \frac{B - \sum_{\tilde{a}\in\mathcal{H}} c_{\tilde{a}}}{\sum_{\tilde{a}\in\{1,\ldots,K\}\backslash[\mathcal{H}\cup\{a^\star\}]} c_{\tilde{a}}}\right\} & \text{if } a \neq K+1 \text{ and KL-UCB.} \end{cases}
$$

Above "Thompson Sampling" and "KL-UCB" in the conditioning statements refers to which of the two algorithms is under consideration. The latter condition represents the extreme scenario where the arms in $\tilde{a} \in \mathcal{H}$ have $I_{\tilde{a}}(t) > \hat{\rho}^\star(t)$, whereas the arms $\tilde{a}$ outside of $\mathcal{H}\cup\{a^\star, K+1\}$ have $I_{\tilde{a}}(t) = \hat{\rho}^\star(t)$. One can verify that $\breve{q}_a^{a^\star} > 0$. Similarly to (24), for each $a \in (\mathcal{M}\cup\mathcal{N})\backslash\{K+1\}$ and $t \geq K$ (KL-UCB) or $t \geq 0$ (Thompson sampling),

$$
\mathbb{P}\left\{I_a(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t), I_{K+1}(t) < \hat{\rho}^\star(t) | \mathcal{F}(t)\right\} \leq \frac{1}{\breve{q}_a^{a^\star}} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), I_{a^\star}(t) < \hat{\rho}^\star(t) \middle| \mathcal{F}(t)\right\},
$$

and thus

$$
\sum_{t=0}^{T-1} \mathbb{P}\left\{I_a(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t), I_{K+1}(t) < \hat{\rho}^\star(t)\right\} \leq \frac{1}{\breve{q}_a^{a^\star}} \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), I_{a^\star}(t) < \hat{\rho}^\star(t)\right\}.
$$

For $a = K+1$, we similarly have

$$
\sum_{t=0}^{T-1} \mathbb{P}\left\{I_{K+1}(t) \geq \hat{\rho}^\star(t), I_{a^\star}(t) < \hat{\rho}^\star(t)\right\} \leq \frac{1}{\breve{q}_a^{a^\star}} \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), I_{a^\star}(t) < \hat{\rho}^\star(t)\right\}.
$$

For each $a \in \mathcal{M}\cup\mathcal{N}$, let

$$
M_a^{a^\star}(T) \equiv \sum_{t=0}^{T-1} \mathbb{1}\{a \in \hat{\mathcal{A}}(t+1), I_{a^\star}(t) < \hat{\rho}^\star(t)\}.
$$

The bounds (28) and (27) yield the key observation that we use in this section:

$$\text{for KL-UCB: } T - K + 1 - \mathbb{E}[N_{a^\star}(T)] \leq \sum_{a \in \mathcal{M} \cup \mathcal{N}} \frac{1}{\check{q}_a^{a^\star}} \mathbb{E}[M_a^{a^\star}(T)];$$
$$\text{for Thompson sampling: } T - \mathbb{E}[N_{a^\star}(T)] \leq \sum_{a \in \mathcal{M} \cup \mathcal{N}} \frac{1}{\check{q}_a^{a^\star}} \mathbb{E}[M_a^{a^\star}(T)]. \tag{29}$$

We note that, for most models $\mathcal{D}_K$, there will generally not be a positive lower bound on $\check{q}_a^{a^\star}$ uniformly over distributions $\mathcal{V}$ in $\mathcal{D}_K$, where we note that the dependence of $\check{q}_a^{a^\star}$ on $\mathcal{V}$ is suppressed in the notation. Therefore, on the one hand, if one were pursuing a worst-case analysis of the regret of our algorithms, where the maximal regret is studied over all $\mathcal{V} \in \mathcal{D}$, then it would typically not be possible to control the right-hand sides above. On the other hand, in our setting, in which we study the regret at a fixed $\mathcal{V}$, it is true that $\check{q}_a^{a^\star} > 0$, and so one can control the right-hand sides above provided they can control $\mathbb{E}[M_a^{a^\star}(T)]$ for arms $a \in \mathcal{M} \cup \mathcal{N}$. In what follows, we will show that we can indeed control $\mathbb{E}[M_a^{a^\star}(T)]$ for these arms.

Let G be some integer in $[0, +\infty($ and $\delta \in (0,1)$ be a constant to be specified shortly. For convenience, we let $T^{(g)} \equiv \lfloor T^{(1-\delta)^g} \rfloor$ for $g \in \mathbb{N}$. We also define

$$\overline{\mathcal{U}} \equiv \{a \in (\mathcal{M} \cup \mathcal{N}) \backslash \{K+1\} : c_a \rho_{a^\star} \geq \mu_+\},$$
$$\underline{\mathcal{U}} \equiv \{a \in (\mathcal{M} \cup \mathcal{N}) \backslash \{K+1\} : c_a \rho_{a^\star} < \mu_+\},$$

where we note that $\overline{\mathcal{U}} \cup \underline{\mathcal{U}} = (\mathcal{M} \cup \mathcal{N}) \backslash \{K+1\}$. Our analysis relies on the following bound (for which we provide the arguments below):

$$\begin{aligned}\sum_{a \in \mathcal{M} \cup \mathcal{N}} \mathbb{E}[M_a^{a^\star}(T)] &\leq \mathbb{E}[N_{K+1}(T)] + \sum_{a \in \overline{\mathcal{U}}} \mathbb{E}[M_a^{a^\star}(T)] + \sum_{a \in \underline{\mathcal{U}}} \mathbb{E}[M_a^{a^\star}(T)] \\ &= o(\log T) + \underbrace{\sum_{a \in \overline{\mathcal{U}}} \mathbb{E}[M_a^{a^\star}(T)] + \sum_{a \in \underline{\mathcal{U}}} \mathbb{E}[M_a^{a^\star}(T^{(G)})]}_{\text{Term A}} \\ &\quad + \underbrace{\sum_{g=1}^{G} \sum_{a \in \underline{\mathcal{U}}} \mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})]}_{\text{Term B}}. \end{aligned} \tag{30}$$

The inequality uses that $\mathbb{E}[M_{K+1}^{a^\star}(T)] \leq \mathbb{E}[N_{K+1}(T)]$, and the equality holds using (i) a telescoping series and (ii) the fact that the algorithm achieves (12): indeed, this was proven for both KL-UCB and Thompson sampling in Section 6.2.

We now present the key ingredients to bound Term A and B. Each lemma stated below holds for both KL-UCB in the settings of Theorems 6 and 7 and for Thompson sampling in the setting of Theorem 8. Though these lemmas hold for both algorithms, the methods of proof for KL-UCB and for Thompson sampling are quite different. Thus we give the proofs of the lemmas in the settings of Theorems 6 and 7 in Appendix C and the proofs in the setting of Theorem 8 in Appendix D.

Lemma 18. *In the settings of Theorem 6, 7, and 8, $\mathbb{E}[M_a^{a^\star}(T)] = o(\log T)$ for $a \in \overline{\mathcal{U}}$ and, for fixed $G \geq 0$,*

$$\mathbb{E}[M_a^{a^\star}(T^{(G)})] \leq (1-\delta)^G \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, c_a \rho_{a^\star})}$$

for $a \in \underline{\mathcal{U}}$. As a consequence,

$$\text{Term A} \leq (1-\delta)^G \sum_{a \in \underline{\mathcal{U}}} \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, c_a \rho_{a^\star})} + o(\log T).$$

The proof of Lemma 18 borrows a lot from the proofs of (11) and (12) for each algorithm.

Controlling Term B relies on a careful choice of $\delta > 0$, which is specified in Lemma 19 below. The proof of this lemma is highly original: indeed we first prove that the considered algorithm is uniformly efficient, which allows to exploit the lower bound (8) given in Theorem 5. Its proof is provided in the appendix for both KL-UCB and Thompson Sampling, and we sketch it below.

Lemma 19. *Let $d \in (0,1)$ and δ chosen such*

$$\delta = d\left[1 - \left(\max_{a \in \mathcal{N} \cap \underline{\mathcal{U}}} \frac{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)}{\mathcal{K}_{\inf}(\nu_a, c_a\rho_{a^\star})}\right)^{1/2}\right], \tag{31}$$

and $\delta = d$ if $\mathcal{N} \cap \underline{\mathcal{U}} = \emptyset$. Then in the setting of Theorems 6, 7, and 8, Term B is $o(\log T)$.

Sketch of proof of Lemma 19. We first show that the algorithms are uniformly efficient in the sense defined in Section 3. This result is an immediate consequence of the results in Section 6.1, which show that the arms in $\mathcal{N}\backslash\{K+1\}$ are not pulled too often, plus the preliminary results in this section, which show that arms in $\mathcal{L}$ are pulled often.

Lemma 20. *KL-UCB is uniformly efficient in the settings of Theorems 6 and 7 and Thompson sampling is uniformly efficient in the setting of Theorem 8.*

Proof. Fix an arbitrary reward distribution $\mathcal{V}$. By by Lemma 18 and the already proven (11) and (12) in the settings of Theorems 6, 7, and 8 (see Lemmas 9, 10, 16, and 17), both of which hold for $\mathcal{V}$,

$$\begin{aligned} T - \mathbb{E}_{\mathcal{V}}[N_{a^\star}(T)] &\leq \sum_{a \in \mathcal{M} \cup \mathcal{N}} \frac{1}{\check{q}_a^{a^\star}} \mathbb{E}_{\mathcal{V}}[M_a^{a^\star}(T)] + O(1) \\ &\leq o(\log T) + \sum_{a \in \overline{\mathcal{U}}} \frac{1}{\check{q}_a^{a^\star}} \mathbb{E}_{\mathcal{V}}[M_a^{a^\star}(T)] + \sum_{a \in \underline{\mathcal{U}}} \frac{1}{\check{q}_a^{a^\star}} \mathbb{E}_{\mathcal{V}}[M_a^{a^\star}(T)] + O(1) \end{aligned}$$

for any $a^\star \in \mathcal{L}$, where the $O(1)$ term is equal to zero for Thompson sampling and, by (29), is $K-1$ for KL-UCB. The right-hand side is $O(\log T)$ by applying the results of Lemma 18 to control the sums over $\overline{\mathcal{U}}$ and $\underline{\mathcal{U}}$. Section 6.1 showed that arms in $\mathcal{N}$ are not pulled often (at most $O(\log T)$ times). By (6), it follows that $R(T) = O(\log T)$, which is $o(T^\alpha)$ for any $\alpha > 0$. □ □

Fix $g \in \mathbb{N}$ and an arm $a \in \mathcal{N} \cap \underline{\mathcal{U}}$. By the uniform efficiency of the algorithm established in Lemma 20, we will be able to apply (8) from Lemma 4 to show that $N_a(T^{(g)}) \geq (1-\delta)\frac{\log T^{(g)}}{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)}$ with probability approaching 1. For now suppose this holds almost surely (in the proofs we deal with the fact that this happens with probability approaching rather than exactly 1). Our objective will be to show that this lower bound on $N_a(T^{(g)})$ suffices to ensure that $M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})$ is $o(\log T)$, in words that arm a is pulled while arm $a^\star$ is pulled with probability zero ($I_{a^\star}(t) < \hat{\rho}^\star(t)$) at most $o(\log T)$ times from time $t = T^{(g)}, \ldots, T^{(g-1)}$.

We will see that $\frac{\log T^{(g-1)}}{\mathcal{K}_{\inf}(\nu_a, c_a\rho_{a^\star})}$ pulls of arm a by time $T^{(g)}$ suffices to ensure this in both settings. Using that $(1-\delta)\log T^{(g)} \approx (1-\delta)^2 \log T^{(g-1)}$, it will follow that we can control the sum in Term B for each $a \in \mathcal{N}$ provided we choose $\delta \in (0,1)$ so that

$$(1-\delta)^2 \frac{1}{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)} > \frac{1}{\mathcal{K}_{\inf}(\nu_a, c_a\rho_{a^\star})} \text{ for all } a \in \mathcal{N}. \tag{32}$$

It is easy to check to for any $d \in (0,1)$, δ as defined in Lemma 19 satisfies this inequality. Note that $\mathcal{K}_{\inf}(\nu_a, c_a\rho_{a^\star}) \geq \mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)$, and thus $\delta \in (0,1)$. So far we have only considered suboptimal arms $a \in \mathcal{N} \cap \underline{\mathcal{U}}$, but the fact that, for any $a \in \mathcal{M} \cap \underline{\mathcal{U}}$, Lemma 4 ensures that $N_a(T^{(g)}) > \log T^{(g)}/\epsilon$ with probability approaching 1 for *any* $\epsilon > 0$ shows that $a \in \mathcal{N} \cap \underline{\mathcal{U}}$ is indeed the harder case. Indeed, this is what we see in our proofs controlling Term B for the two algorithms. □ □

We now conclude the analysis. Combining Equations (29) and (30) with the bounds on Term A and B obtained in Lemma 18 and Lemma 19 yield, for any finite G and for the particular choice of $\delta \in (0,1)$ given in (31)

$$\limsup_T \frac{T - \mathbb{E}[N_{a^\star}(T)]}{\log T} \leq (1-\delta)^G \sum_{a \in \underline{\mathcal{U}}} \frac{1}{\check{q}_a^{a^\star} \, \mathcal{K}_{\inf}(\nu_a, c_a \rho_{a^\star})}. \tag{33}$$

Taking G to infinity yields the result.

7 Conclusion

We have established the asymptotic efficiency of KL-UCB and Thompson sampling for budgeted multiple-play bandit problem in which the cost of pulling each arm is known and, in each round, the agent may use any strategy for which the expected cost is no more than their budget. We have also introduced a pseudo-arm so that the agent has the option of reserving the remainder of their budget if the remaining arms have reward-to-cost ratios that fall below a prespecified indifference point. Thompson sampling outperforms KL-UCB in three of our four simulations scenarios. Despite the strong performance of Thompson sampling for Bernoulli rewards, we have been able to prove stronger results about KL-UCB in this work, dealing with more general distributions. Understanding for which distributions one of these algorithms is preferable to the other is an interesting area for future work.

All of the proofs in this work can handle the case that the set of optimal arms is not unique. In an earlier work, Komiyama et al. [2015] established the optimality of Thompson sampling under a multiple play bandit model in which the set of optimal arms is unique. A potential area for future work would be to extend their arguments to the special case of our budgeted bandit setting in which the set of optimal arms is unique – it would be interesting to see if their technique yields a shorter proof in this special case.

In future work, it would be interesting to consider an extension of our setting where the budget (B_t), indifference points (ρ_t), and costs (c_t) are random over time according to some exogeneous source of randomness. If only the budget is random over time, then, under some regularity conditions, the regret lower bound and regret of our algorithms would seem to be driven by the behavior of our algorithm for the fixed budget representing the upper edge of the support for the random budget, since this is the setting in which the most information is learned about the arm distributions (arms that are otherwise suboptimal can be optimal in this setting). If only the indifference point is variable over time, then the behavior of our algorithm will similarly be driven by the lowest indifference point, since the most information is available in this case. Combinations of variable budgets and indifference points will result in a similar analysis. Variable but known costs are more complex, because they have the potential to change the order and indices of the optimal arms. For sufficiently variable costs, we in fact expect that all arms will be pulled more than order $\log T$ times, since all arms will be optimal for certain cost realizations. Therefore, a careful study of a variable cost budgeted bandit problem may require very different techniques than those used in this work.

Structure of the Supplementary Material Appendix A focuses on oracle strategy and regret decomposition. Appendix B contains proofs establishing the asymptotic lower bound on the number of suboptimal arm draws. Appendices C and D contain technical proofs for KL-UCB and the Thompson sampling, respectively.

Acknowledgements The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-13-BS01-0005 (project SPADRO) and ANR-16-CE40-0002 (project BADASS). Alex Luedtke gratefully acknowledges the support of a Berkeley Fellowship.

References

S Agrawal and N R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006. ACM, 2014.

S Agrawal and N Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *arXiv preprint arXiv:1111.1797*, 2011.

S Agrawal and N Goyal. Further optimal regret bounds for thompson sampling. *arXiv preprint arXiv:1209.3353*, 2012.

V Anantharam, P Varaiya, and J Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards. *Automatic Control, IEEE Transactions on*, 32(11):968–976, 1987.

J-Y Audibert, S Bubeck, and G Lugosi. Minimax policies for combinatorial prediction games. *arXiv preprint arXiv:1105.4871*, 2011.

A Badanidiyuru, R Kleinberg, and A Slivkins. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 207–216. IEEE, 2013.

A N Burnetas and M Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

O Cappé, A Garivier, O A Maillard, R Munos, and G Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013a.

O Cappé, A Garivier, O A Maillard, R Munos, and G Stoltz. Supplement to "KullbackLeibler upper confidence bounds for optimal sequential allocation.". *doi:10.1214/13-AOS1119SUPP*, 2013b.

N Cesa-Bianchi and G Lugosi. Combinatorial Bandits. *Journal of Computer and System Sciences*, 78:1404–1422, 2012.

W Chen, Y Wang, and Y Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

R Combes, S Magureanu, A Proutière, and C Laroche. Learning to Rank: Regret Lower Bounds and Efficient Algorithms. In *Proceedings of the 2015 {ACM} {SIGMETRICS} International Conference on Measurement and Modeling of Computer Systems*, pages 231–244, 2015a.

R Combes, M S T M Shahi, A Proutiere, and M Lelarge. Combinatorial Bandits Revisited. In *Advances in Neural Information Processing Systems*, pages 2107–2115, 2015b.

G B Dantzig. Discrete-variable extremum problems. *Operations research*, 5(2):266–288, 1957.

A Garivier, P Ménard, and G Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. *arXiv preprint arXiv:1602.07182*, 2016.

J C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2): 148–177, 1979.

R M Karp. *Reducibility among combinatorial problems*. Springer, New York Berlin Heidelberg, 1972.

E Kaufmann, O Cappé, and A Garivier. On Bayesian upper confidence bounds for bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pages 592–600, 2012a.

E Kaufmann, N Korda, and R Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213. Springer, 2012b.

J Komiyama, J Honda, and H Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. *arXiv preprint arXiv:1506.00779*, 2015.

N Korda, E Kaufmann, and R Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

B Kveton, Z Weng, A Ashkan, E Hoda, and B Eriksson. Matroid Bandits: Fast Combinatorial Optimization with Learning. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

B Kveton, C Szepesvári, Z Wen, and A Ashkan. Cascading Bandits: Learning to Rank in the Cascade Model. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 767–776, 2015a.

B Kveton, W Zheng, A Ashkan, and C Szepesvári. Combinatorial Cascading Bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2015b.

P Lagrée, C Vernade, and O Cappé. Multiple-Play Bandits in the Postition-Based Model. *Preprint, arXiv:1606.02448*, 2016.

T L Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

T L Lai and H Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

H Li and Y Xia. Infinitely Many-Armed Bandits with Budget Constraints. In *AAAI*, pages 2182–2188, 2017.

A Luedtke, E Kaufmann, and A Chambaz. Asymptotically Optimal Algorithms for Multiple Play Bandits with Partial Feedback. *arXiv preprint arXiv:1606.09388*, 2016.

H Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5): 527–535, 1952.

K Sankararaman and A Slivkins. Combinatorial semi-bandits with knapsacks. In *AISTATS*, 2018.

W R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

L Tran-Thanh, A C Chapman, A Rogers, and N R Jennings. Knapsack Based Optimal Policies for Budget-Limited Multi-Armed Bandits. In *AAAI*, 2012.

Z Wen, B Kveton, and A Ashkan. Efficient Learning in Large-Scale Combinatorial Semi-Bandits. In *International Conference on Machine Learning (ICML)*, 2015.

Y Xia, H Li, T Qin, N Yu, and T-Y Liu. Thompson Sampling for Budgeted Multi-Armed Bandits. In *IJCAI*, pages 3960–3966, 2015.

Y Xia, W Ding, X-D Zhang, N Yu, and T Qin. Budgeted bandit problems with continuous random costs. In *Asian Conference on Machine Learning*, pages 317–332, 2016a.

Y Xia, T Qin, W Ma, N Yu, and T-Y Liu. Budgeted Multi-Armed Bandits with Multiple Plays. In *IJCAI*, pages 2210–2216, 2016b.

Appendix

We begin with an outline of the results proven in this appendix and how they are related to one another. Lemma 4 gives a lower bound on the number of draws of each suboptimal arm for a uniformly efficient algorithm. Deduced from Lemma 4, Theorem 5 gives an asymptotic regret lower bound (10) for a uniformly efficient algorithm. The asymptotic lower bound is achieved whenever the expected number of draws of each suboptimal arm satisfies the appropriate asymptotic condition, either (11) or (12) depending on the arm, and the expected number of draws of each optimal arm away from the margin satisfies the asymptotic condition (13). Theorems 6 and 7 state that the variants of KL-UCB are uniformly efficient and achieve (10) for rewards sampled either from a single parameter exponential family or from bounded and finitely supported distributions. Theorem 8 states that Thompson sampling is uniformly efficient and achieves (10) for Bernoulli distributed rewards.

The first step of the proof of Theorems 6, 7, and 8 consists in showing that KL-UCB and Thompson sampling achieve the asymptotically optimal expected number of suboptimal arm draws, i.e. that (11) and (12) hold in their contexts. For KL-UCB, this is a consequence of a preliminary analysis given in Lemmas A.1 and A.2. For Thompson sampling, this is a consequence of another preliminary analysis given in Lemmas 11 through 14. The proof of Lemma 14 relies on a link between the beta and binomial distributions given in Lemma A.3.

The second step of the proof of Theorems 6, 7, and 8 consists in showing that KL-UCB and Thompson sampling are uniformly efficient in their respective contexts. This is a consequence of yet another preliminary analysis, (11), (12), and Lemma 18.

The third step of the proof of Theorems 6, 7, and 8 consists in showing that KL-UCB and Thompson sampling achieve the asymptotically optimal expected number of optimal draws away from the margin, i.e. that (13) holds in their contexts. This is a consequence of the preliminary analysis undertaken in step two and of Lemmas 18 and 19. The proofs of Lemmas 18 and 19 hinge on Lemmas 11 through 14. The proof of Lemma 19 also relies on Lemma A.3.

The fourth and final step of the proof of Theorems 6, 7, and 8 boils down to applying Theorem 5.

A Oracle strategy and regret decomposition

A.1 Proof of Proposition 1

Recall that

$$\boldsymbol{q}^\star \in \operatorname*{argmax}_{\boldsymbol{q}\in[0,1]^K} \sum_{a=1}^K q_a(\mu_a - c_a\rho) \quad \text{such that} \quad \sum_{a=1}^K q_a c_a \leq B.$$

Introducing $c_{K+1} = B$ and $\mu_{K+1} = B\rho$, one can prove that $\boldsymbol{q}^\star$ coincides with the first K components of $\boldsymbol{q}^\star_{K+1} \in [0,1]^{K+1}$, that is defined as the solution to

$$\boldsymbol{q}^\star_{K+1} \in \operatorname*{argmax}_{\boldsymbol{q}\in[0,1]^{K+1}} \sum_{a=1}^{K+1} q_a(\mu_a - c_a\rho) \quad \text{such that} \quad \sum_{a=1}^{K+1} q_a c_a = B \tag{A.1}$$

and that the two optimization problems have the same value. This is because as $\mu_{K+1} - c_{K+1}\rho = 0$, the two objective functions coincide:

$$f_K(\boldsymbol{q}) \equiv \sum_{a=1}^K q_a(\mu_a - c_a\rho) = \sum_{a=1}^{K+1} q_a(\mu_a - c_a\rho) \equiv f_{K+1}(\boldsymbol{q}_{K+1})$$

and if $\boldsymbol{q}$ satisfies the first constraint, there exists q_{K+1} such that $\boldsymbol{q}_{K+1} = (\boldsymbol{q}, q_{K+1})$ satisfies the second constraint: $\sum_{a=1}^{K+1} q_a c_a = B$ (as $c_{K+1} = B$). Conversely, if $\boldsymbol{q}_{K+1}$ satisfies the second constraint, its first K marginals clearly satisfy the first constraint.

The common value $M^\star$ of these two optimization problem, that is the maximal achievable reward, can be rearranged a bit, using that $\sum_{a=1}^{K+1} q_a c_a \rho = \rho B$:

$$M^\star = \sum_{a=1}^{K+1} q_a^\star \mu_a - \rho B,$$

where $\boldsymbol{q}_{K+1}^\star \in [0,1]^{K+1}$ is the solution to

$$\boldsymbol{q}_{K+1}^\star \in \operatorname*{argmax}_{\boldsymbol{q}\in[0,1]^{K+1}} \sum_{a=1}^{K+1} q_a \mu_a \quad \text{such that} \quad \sum_{a=1}^{K+1} q_a c_a = B. \tag{A.2}$$

Now introduce

$$L^\star \equiv \sum_{a\in\mathcal{L}} \mu_a + \rho^\star \left(B - \sum_{a\in\mathcal{L}} c_a \right).$$

The optimal weights are also defined by

$$\boldsymbol{q}_{K+1}^\star \in \operatorname{argmin} \boldsymbol{q} \in [0,1]^{K+1} \left[L^\star - \sum_{a=1}^{K+1} q_a \mu_a \right] \quad \text{such that} \quad \sum_{a=1}^{K+1} q_a c_a = B.$$

The new objective can be rewritten as follows, where the 'virtual' arm $K+1$ that has characteristics $\mu_{K+1} = B\rho$ and $c_{K+1} = B$ is added to either the set $\mathcal{M}$ (if $\rho_{K+1} = \rho = \rho^\star$) or $\mathcal{N}$ (if $\rho_{K+1} = \rho < \rho^\star$).

$$\begin{aligned}
L^\star - \sum_{a=1}^{K+1} q_a \mu_a &= \sum_{a\in\mathcal{L}} \mu_a + \rho^\star \left(B - \sum_{a\in\mathcal{L}} c_a \right) - \sum_{a\in\mathcal{L}} c_a q_a \rho_a - \sum_{a\in\mathcal{M}} c_a q_a \rho^\star - \sum_{b\in\mathcal{N}} c_b q_b \rho_b \\
&= \sum_{a\in\mathcal{L}} c_a \rho_a + \rho^\star \left(B - \sum_{a\in\mathcal{L}} c_a \right) - \sum_{a\in\mathcal{L}} c_a q_a \rho_a - \rho^\star \left(B - \sum_{a\in\mathcal{L}} c_a q_a - \sum_{b\in\mathcal{N}} c_b q_b \right) - \sum_{b\in\mathcal{N}} c_b q_b \rho_b \\
&= \sum_{a\in\mathcal{L}} c_a \underbrace{(\rho_a - \rho^\star)}_{>0} (1 - q_a) + \sum_{b\in\mathcal{N}} c_b \underbrace{(\rho^\star - \rho_b)}_{>0} q_b.
\end{aligned}$$

This shows that the objective function is always non negative, and that it can actually be set to the zero by choosing weights that satisfy $q_a = 1$ for all $a \in \mathcal{L}$ and $q_b = 0$ for all $b \in \mathcal{L}$.

It remains to justify that such a choice is indeed feasible for some choices of weights on the arms in the margin $\mathcal{M}$. This margin is never empty, as in the case $\rho^\star = \rho$, it does contain the 'pseudo-arm' mentioned above. By definition of the sets $\mathcal{L}$ and $\mathcal{M}$,

$$\sum_{a\in\mathcal{L}} c_a < B \quad \text{and} \quad \sum_{a\in\mathcal{L}\cup\mathcal{M}} c_a \geq B$$

hence, the solution can be "completed" by putting weight on the margin such that $\sum_{a\in\mathcal{L}} c_a + \sum_{a\in\mathcal{N}} q_a c_a = B$.

If $\rho < \rho^\star$, then the arm $K+1$ belongs to $\mathcal{N}$ and as such $q_{K+1} = 0$ and the first K marginals indeed satisfy the statement of Proposition 1, with a non-empty margin. If $\rho = \rho^\star$, our 'extended' margin only contains arm $K+1$, while the original margin is empty. As such the only arms with non-zero weights among the first K marginals are the arms in $\mathcal{L}$, for which the weight is one.

A.2 Proof of Proposition 3

$$\text{Regret}(T, \mathcal{V}) = \mathbb{E}\left[\sum_{t=1}^{T} (G^\star - G(t)) \right] = \mathbb{E}\left[\sum_{t=1}^{T} (G^\star - \sum_{a=1}^{K} q_a(t)(\mu_a - c_a \rho)) \right]$$

The proof follows from a rewriting of

$$G^\star - \sum_{a=1}^{K} q_a(t)(\mu_a - c_a\rho) = \sum_{a\in\mathcal{L}} c_a\rho_a + \rho^\star\left(B - \sum_{a\in\mathcal{L}} c_a\right) - B\rho - \sum_{a=1}^{K} q_a(t)(\mu_a - c_a\rho)$$
$$= \sum_{a\in\mathcal{L}} c_a\rho_a + \rho^\star\left(B - \sum_{a\in\mathcal{L}} c_a\right) - B\rho - \sum_{a=1}^{K+1} q_a(t)(\mu_a - c_a\rho),$$

where we define $\mu_{K+1} = \rho B$, $c_K = B$ and let $q_{K+1}(t)$ be such that $\sum_{a=1}^{K+1} q_a(t)c_a = B$. This is possible as $\sum_{a=1}^{K} q_a(t)c_a \leq B$ due to the soft budget constraints and $c_{K+1} = B$ and

$$q_{K+1}(t) = \frac{B - \sum_{a=1}^{K} c_a q_a(t)}{B}.$$

Thus one can further write

$$G^\star - \sum_{a=1}^{K} q_a(t)(\mu_a - c_a\rho) = \sum_{a\in\mathcal{L}} c_a\rho_a + \rho^\star\left(B - \sum_{a\in\mathcal{L}} c_a\right) - \sum_{a=1}^{K+1} q_a(t)\mu_a$$
$$= \sum_{a\in\mathcal{L}} c_a\rho_a + \rho^\star\left(B - \sum_{a\in\mathcal{L}} c_a\right) - \sum_{a\in\mathcal{L}} q_a(t)c_a\rho_a - \rho^\star\sum_{a\in\mathcal{M}} q_a(t)c_a - \sum_{a\in\mathcal{N}} q_a(t)c_a\rho_a - q_{K+1}(t)\rho B$$

Using that

$$\sum_{a\in\mathcal{M}} q_a(t)c_a = B - \sum_{a\in\mathcal{L}} q_a(t)c_a - \sum_{a\in\mathcal{M}} q_a(t)c_a - q_{K+1}(t)B,$$

one obtains

$$G^\star - \sum_{a=1}^{K} q_a(t)(\mu_a - c_a\rho) = \sum_{a\in\mathcal{L}} c_a(\rho_a - \rho^\star)(1 - q_a(t)) + \sum_{a\in\mathcal{N}} c_a(\rho^\star - \rho_a)q_a(t) + B(\rho^\star - \rho)q_{K+1}(t)$$
$$= \sum_{a\in\mathcal{L}} c_a(\rho_a - \rho^\star)(1 - q_a(t)) + \sum_{a\in\mathcal{N}} c_a(\rho^\star - \rho_a)q_a(t) + (\rho^\star - \rho)\left(B - \sum_{a=1}^{K} c_a q_a(t)\right)$$

Summing over t, the regret can be decomposed as

$$\sum_{a\in\mathcal{L}} c_a(\rho_a - \rho^\star)\left(T - \mathbb{E}\left[\sum_{t=1}^{T} q_a(t)\right]\right) + \sum_{a\in\mathcal{N}} c_a(\rho^\star - \rho_a)\mathbb{E}\left[\sum_{t=1}^{T} q_a(t)\right] + (\rho^\star - \rho)\left(B - \sum_{a=1}^{K} c_a\mathbb{E}\left[\sum_{t=1}^{T} q_a(t)\right]\right)$$

and the conclusion follows by noting that $N_a(T) = \mathbb{E}\left[\sum_{a=1}^{T} q_a(t)\right]$.

B Proof of Lower Bound on Suboptimal Arm Draws

Proof of Lemma 4. Fix some arm $a \in (\mathcal{M}\cup\underline{\mathcal{N}})\backslash\{K+1\}$, natural number T, and $\delta \in (0,1)$. By definition, $c_a\rho^\star < \mu_+$ for all $a \in \underline{\mathcal{N}}$, and, for $a \in \mathcal{M}$ the same property holds by our assumption that $c_a\rho^\star = \mu_a < \mu_+$. Hence, the set $\{\tilde{\nu}_a \in \mathcal{D} : E(\tilde{\nu}_a) > c_a\rho^\star\}$ is non-empty. If the intersection of this set with the set of distributions $\{\tilde{\nu}_a \in \mathcal{D} : \nu_a \ll \tilde{\nu}_a\}$ is empty, then the bounds are trivial by our convention that $d/\infty = 0$ for finite d. Otherwise, let $\mathcal{V}'$ be some distribution that is equal to $\mathcal{V}$ except in the a^{th} component, where its a^{th} component $\nu'_a \in \mathcal{D}$ is such that $\mu'_a \equiv E(\nu'_a) > c_a\rho^\star$ and $\nu_a \ll \nu'_a$. Furthermore, one can select $\mathcal{V}'$ to fall in the statistical model for the joint distribution of the arm-specific rewards by our variation-independence assumption. For each b, let $\rho'_b = \rho_b$, $b \neq a$, and let $\rho'_a = \mu'_a/c_a$. Observe

that $\mu'_a > c_a\rho^\star \geq \mu_a$ implies that $\mathrm{KL}(\nu_a, \nu'_a) > 0$. Define the log-likelihood ratio random variable $L_a(T) \equiv L_{a,N_a(T)} \equiv \sum_{n=1}^{N_a(T)} \log \frac{d\nu_a}{d\nu'_a}(X_{a,n})$. Let $b_a(T) \equiv (1-\delta)\frac{\log T}{\mathrm{KL}(\nu_a,\nu'_a)}$ and $d(T) \equiv (1-\delta/2)\log T$. We have that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{V}}\left\{N_a(T) < b_a(T)\right\} \\
&\leq \mathbb{P}_{\mathcal{V}}\left\{N_a(T) < b_a(T), L_a(T) \leq d(T)\right\} + \mathbb{P}_{\mathcal{V}}\left\{N_a(T) < b_a(T), L_a(T) > d(T)\right\} \\
&\leq e^{d(T)}\,\mathbb{P}_{\mathcal{V}'}\left\{N_a(T) < b_a(T)\right\} + \mathbb{P}_{\mathcal{V}}\left\{N_a(T) < b_a(T), L_a(T) > d(T)\right\},
\end{aligned} \tag{A.3}$$

where the final inequality holds because, for any event $D \subseteq \{N_a(T) = b, L_a(T) \leq d(T)\}$, a change of measure shows that $\mathbb{P}_{\mathcal{V}}\{D\} = \mathbb{E}_{\mathcal{V}'}\left[e^{L_{a,b}}\mathbb{1}_{\{D\}}\right] \leq e^{d(T)}\,\mathbb{P}_{\mathcal{V}'}\{D\}$ [see Equation 2.6 in Lai and Robbins, 1985]. Let $\tilde{\rho}^\star \equiv \rho^\star(c_a\rho'_a : a = 1,\ldots,K+1)$. Observe that arm a under the reward distribution involving ν'_a satisfies either (i) $\rho'_{a'} > \tilde{\rho}^\star$ or (ii) $\rho'_a = \tilde{\rho}^\star$ and $g_a \equiv B - \sum_{\tilde{a}\neq a:\rho_{\tilde{a}}\geq\tilde{\rho}^\star} c_{\tilde{a}} > 0$, where the sum over the empty set is zero. Under (i), we note that the uniform efficiency of the algorithm and Markov's inequality yield that

$$\mathbb{P}_{\mathcal{V}'}\left\{N_a(T) < b_a(T)\right\} = \mathbb{P}_{\mathcal{V}'}\left\{T - N_a(T) > T - b_a(T)\right\} = o\left(T^{\delta/2-1}\right).$$

Thus, the first term in (A.3) converges to zero as $T \to \infty$ when (i) holds. We now show the same result when (ii) holds. We first note that

$$g_aT - c_a\,\mathbb{E}[N_a(T)] \geq BT - \sum_{\tilde{a}\neq a:\rho_{\tilde{a}}\geq\tilde{\rho}^\star} c_{\tilde{a}}\,\mathbb{E}[N_{\tilde{a}}(T)] - c_a\,\mathbb{E}[N_a(T)] = BT - \sum_{\tilde{a}:\rho_{\tilde{a}}\geq\tilde{\rho}^\star} c_{\tilde{a}}\,\mathbb{E}[N_{\tilde{a}}(T)].$$

The right-hand side is $o(T^{\delta/2})$ by the uniform efficiency of the algorithm. Hence, Markov's inequality yields that,

$$\mathbb{P}_{\mathcal{V}'}\left\{N_a(T) < b_a(T)\right\} = \mathbb{P}_{\mathcal{V}'}\left\{g_aT - c_aN_a(T) > g_aT - c_ab_a(T)\right\} = o\left(T^{\delta/2-1}\right).$$

Thus, the first term in (A.3) also converges to zero as $T \to \infty$ when (ii) holds. For the second term, observe that

$$\begin{aligned}
\{N_a(T) < b_a(T), L_a(T) > d(T)\} &\subseteq \left\{\max_{n\leq b_a(T)} \frac{L_{a,n}}{b_a(T)} > \frac{d(T)}{b_a(T)}\right\} \\
&= \left\{\max_{n\leq b_a(T)} \frac{L_{a,n}}{b_a(T)} > \frac{1-\delta/2}{1-\delta}\,\mathrm{KL}(\nu_a,\nu'_a) > \mathrm{KL}(\nu_a,\nu'_a)\right\}.
\end{aligned}$$

By the strong law of large numbers, $b_a(T)^{-1}L_{a,\lfloor b_a(T)\rfloor} \to \mathrm{KL}(\nu_a,\nu'_a)$ almost surely under ν_a. Further, $\max_{n\leq b_a(T)} b_a(T)^{-1}L_{a,n} \to \mathrm{KL}(\nu_a,\nu'_a)$ almost surely as $T \to \infty$. It follows that the second term in (A.3) converges to zero as $T \to \infty$ so that

$$\mathbb{P}_{\mathcal{V}}\left\{N_a(T) < (1-\delta)\frac{\log(T)}{\mathrm{KL}(\nu_a,\nu'_a)}\right\} \to 0. \tag{A.4}$$

For convenience, we let $\mathcal{K} \equiv \mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)$ in what follows. By the definition of the infimum, for every $\epsilon > 0$ there exists some ν'_a such that $\mathcal{K} + \epsilon > \mathrm{KL}(\nu_a,\nu'_a)$. This proves (8). If $a \in \underline{\mathcal{N}}$ so that $\mathcal{K} > 0$, then take $\epsilon = \left[(1-\delta)^{-1/2} - 1\right]\mathcal{K}$ and write

$$\mathbb{P}_{\mathcal{V}}\left\{N_a(T) < (1-\delta)^{3/2}\frac{\log(T)}{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)}\right\} \to 0.$$

Applying the above to $\delta' = 1-(1-\delta)^{2/3}$ (such that $(1-\delta')^{3/2} = (1-\delta)$) yield the result for $a \in \underline{\mathcal{N}}$. For $a \in \underline{\mathcal{N}}$, it also follows that for all $\delta \in (0,1)$ one has

$$\mathbb{E}[N_a(T)] \geq \frac{(1-\delta)\log T}{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)}\,\mathbb{P}_{\mathcal{V}}\left\{N_a(T) \geq (1-\delta)\frac{\log T}{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)}\right\} \underset{T\to\infty}{\sim} \frac{(1-\delta)\log T}{\mathcal{K}_{\inf}(\nu_a, c_a\rho^\star)},$$

which yields (9), letting δ go to zero. □ □

C Supplementary Proofs for KL-UCB

Lemma A.1. *Fix an $a \in \{1, \dots, K\}$ and a fixed $\mu^\dagger$ (not relying on T) with $\mu_a < \mu^\dagger$. In the setting of Theorem 6 with $\rho^\dagger = \mu^\dagger / c_a$ or in the setting of Theorem 7 with $\rho^\dagger = \left[1 - \log(T)^{-1/5}\right] \mu^\dagger / c_a$, it holds that*

$$\sum_{n=b(T)+1}^{\infty} \mathbb{P}\left\{\hat{\nu}_{a,n} \in \mathcal{C}_{c_a \rho^\dagger, f(T)/n}\right\} = o(\log T),$$

where $b(T)$ is any number satisfying

$$b(T) \geq \left\lceil \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger)} \right\rceil.$$

An explicit finite sample bound on the $o(\log T)$ term can be found in Cappé et al. [2013b].

Proof. In the setting of Theorem 6, Equation 25 in Cappé et al. [2013b] gives the result for $\rho^\dagger = \mu^\dagger / c_a$. We refer the readers to that equation for the explicit finite sample bound that we are summarizing with little-oh notation.

In the setting of Theorem 7, Equation 33 combined with the unnumbered equation preceding Equation 36 in Section B.4 of Cappé et al. [2013b] gives the result for $\rho^\dagger = \left[1 - \log(T)^{-1/5}\right] \mu^\dagger / c_a$. An explicit finite sample upper bound on this quantity can be found in Section B.4 of Cappé et al. [2013b]. □ □

Lemma A.2. *Fix an arm $a^\star \in \mathcal{S}$. In the setting of Theorem 6 with $\rho^\dagger \leq \rho_{a^\star}$ or in the setting of Theorem 7 with $\rho^\dagger \leq \left[1 - \log(T)^{-1/5}\right] \rho_{a^\star}$, it holds that*

$$\sum_{t=K}^{T-1} \mathbb{P}\left\{c_{a^\star} \rho^\dagger \geq U_{a^\star}(t)\right\} = o(\log T).$$

Explicit finite sample constants can be found in the proof.

Proof. In the setting of Theorem 6, it holds that $\{c_{a^\star} \rho^\dagger \geq U_{a^\star}(t)\} \subseteq \{\mu_{a^\star} \geq U_{a^\star}(t)\}$. Hence,

$$\sum_{t=K}^{T-1} \mathbb{P}\left\{c_{a^\star} \rho^\dagger \geq U_{a^\star}(t)\right\} \leq \sum_{t=K}^{T-1} \mathbb{P}\{\mu_{a^\star} \geq U_{a^\star}(t)\}.$$

Furthermore,

$$\{\mu_{a^\star} \geq U_{a^\star}(t)\} \subseteq \bigcup_{n=1}^{t-K+1} \left\{\mu_{a^\star} \geq \hat{\mu}_{a^\star,n}, \mathrm{KL}(\hat{\mu}_{a^\star,n}, \mu_{a^\star}) \geq \frac{f(t)}{n}\right\}.$$

Using the above, Equations 17 and 18 in Cappé et al. [2013b] show that $\sum_{t=K}^{T-1} \mathbb{P}\{\mu_{a^\star} \geq U_{a^\star}(t)\}$ is upper bounded by $3 + 4e \log\log T = o(\log T)$ provided $T \geq 3$.

In the setting of Theorem 7, it holds that $\{c_{a^\star} \rho^\dagger \geq U_{a^\star}(t)\} \subseteq \{\left[1 - \log(T)^{-1/5}\right] \mu_{a^\star} \geq U_{a^\star}(t)\}$. Hence,

$$\sum_{t=K}^{T-1} \mathbb{P}\left\{c_{a^\star} \rho^\dagger \geq U_{a^\star}(t)\right\} \leq \sum_{t=K}^{T-1} \mathbb{P}\left\{\left[1 - \log(T)^{-1/5}\right] \mu_{a^\star} \geq U_{a^\star}(t)\right\}. \tag{A.5}$$

Let $\epsilon \equiv \log(T)^{-1/5} \mu_{a^\star} > 0$. Arguments given in Section B.2 of Cappé et al. [2013b] show that

$$\begin{aligned}
\{\mu_{a^\star} - \epsilon \geq U_{a^\star}(t)\} &\subseteq \left\{\mathcal{K}_{\inf}(\hat{\nu}_{a^\star}(t), \mu_{a^\star} - \epsilon) \geq \frac{f(t)}{N_{a^\star}(t)}\right\} \\
&\subseteq \left\{\mathcal{K}_{\inf}(\hat{\nu}_{a^\star}(t), \mu_{a^\star}) \geq \frac{f(t)}{N_{a^\star}(t)} + \frac{\epsilon^2}{2}\right\}
\end{aligned}$$

$$\subseteq \cup_{n=1}^{t-K+1} \left\{ \mathcal{K}_{\inf}(\hat{\nu}_{a^\star,n}, \mu_{a^\star}) \geq \frac{f(t)}{n} + \frac{\epsilon^2}{2} \right\}.$$

The remainder of the proof is now the same as in Cappé et al. [2013b]. In particular, their Equation 26 combined with the bounds given after their Equation 35 shows that the right-hand side of (A.5) is upper bounded by $36\mu_{a^\star}^{-4}\left(2+\log\log T\right)(\log T)^{4/5} = o(\log T)$. □ □

Proof of Lemma 18 for KL-UCB in the settings of Theorems 6 and 7. Fix $a \in \underline{\mathcal{U}} \cup \overline{\mathcal{U}}$. For ease of notation, we analyze $\mathbb{E}[M_a^{a^\star}(T)]$ rather than $\mathbb{E}[M_a^{a^\star}(T^{(G)})]$, but for fixed $G < \infty$ there is no loss of generality in doing so. If $a \in \underline{\mathcal{U}}$, then let $\mu^\dagger = c_a \rho_{a^\star}$, and otherwise, fix $\mu^\dagger \in (\mu_a, \mu_+)$. Let $\rho^\dagger \equiv \mu^\dagger / c_a$ (setting of Theorem 6) or $\rho^\dagger \equiv \left[1 - \log(T)^{-1/5}\right] \mu^\dagger / c_a$ (setting of Theorem 7). Note that $\rho^\dagger < \mu_+ / c_a$. Analogous arguments to those used for (20) show that

$$\begin{aligned}
&\left\{ a \in \hat{\mathcal{A}}(t+1), \frac{U_{a^\star}(t)}{c_{a^\star}} < \hat{\rho}^\star(t) \right\} \\
&\subseteq \left\{ a \in \hat{\mathcal{A}}(t+1), \frac{U_{a^\star}(t)}{c_{a^\star}} < \hat{\rho}^\star(t), \rho^\dagger \geq \frac{U_a(t)}{c_a} \right\} \cup \left\{ a \in \hat{\mathcal{A}}(t+1), \frac{U_{a^\star}(t)}{c_{a^\star}} < \hat{\rho}^\star(t), \rho^\dagger < \frac{U_a(t)}{c_a} \right\} \\
&\subseteq \left\{ \rho^\dagger \geq \frac{U_{a^\star}(t)}{c_{a^\star}} \right\} \cup \left\{ a \in \hat{\mathcal{A}}(t+1), \rho^\dagger < \frac{U_a(t)}{c_a} \right\}. \qquad \text{(A.6)}
\end{aligned}$$

Let

$$b_a^{a^\star}(T) \equiv \left\lceil \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger)} \right\rceil.$$

Similarly to (22), we have that

$$\mathbb{E}[M_a^{a^\star}(T)] \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu_{a^\star})} + \sum_{n=b_a^{a^\star}(T)+1}^{\infty} \mathbb{P}\left\{ \hat{\nu}_{a,n} \in \mathcal{C}_{c_a\rho^\dagger, f(T)/n} \right\} + \sum_{t=K}^{T-1} \mathbb{P}\left\{ \rho^\dagger \geq \frac{U_{a^\star}(t)}{c_{a^\star}} \right\} + 2.$$

By Lemmas A.1 and A.2,

$$\mathbb{E}[M_a^{a^\star}(T)] \leq \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger)} + o(\log T). \qquad \text{(A.7)}$$

In what follows we refer to this $o(\log T)$ term as $r(T, \mu^\dagger)$, where we note that $r(T, \mu^\dagger)/\log T \to 0$ for each fixed $\mu^\dagger \in (\mu_a, \mu_+)$. If $a \in \overline{\mathcal{U}}$, we will obtain our result by letting $\mu^\dagger \to \mu_+$. Thus, there exists a sequence $\mu^\dagger(T) \to \mu_+$ such that $r(T, \mu^\dagger(T))/\log T \to 0$. Noting $\liminf_{\mu^\dagger \to \mu_+} \mathcal{K}_{\inf}(\nu_a, \mu^\dagger) = +\infty$ in the setting of both theorems, we see that

$$\mathbb{E}[M_a^{a^\star}(T)] \leq \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger(T))} + r(T, \mu^\dagger(T)) = o(\log T).$$

This is the desired result when $a \in \overline{\mathcal{U}}$. If, instead, $a \in \underline{\mathcal{U}}$, then replacing T by $T^{(G)}$ in (A.7) (for T large enough so that $T^{(G)} > 1$), and recalling that $\mu^\dagger = c_a \rho_{a^\star}$ when $a \in \underline{\mathcal{U}}$, gives the desired result. □ □

Proof of Lemma 19 for KL-UCB in the settings of Theorems 6 and 7. Fix $g \in \mathbb{N}$, $a \in \underline{\mathcal{U}} \subset \mathcal{M} \cup \mathcal{N}$, and $T^{(g)}$ such that $T^{(g)} > 1$. In the setting of Theorem 6 let $\rho^\dagger = \rho_{a^\star}$, and in the setting of Theorem 7 let $\rho^\dagger = [1 - \log(T)^{-1/5}]\rho_{a^\star}$. By (A.6) and the fact that $\{\rho^\dagger < U_a(t)/c_a\} = \left\{\hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{c_a\rho^\dagger, f(t)/N_a(t)}\right\}$,

$$\begin{aligned}
&\mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})] \\
&\leq \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{ \rho^\dagger \geq \frac{U_{a^\star}(t)}{c_{a^\star}} \right\} + \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{ a \in \hat{\mathcal{A}}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{c_a\rho^\dagger, f(t)/N_a(t)} \right\}.
\end{aligned}$$

The first term in the right hand side is upper bounded by the same sum from $t = K$ to $T - 1$, and is thus $o(\log T)$ by Lemma A.2. For the second term, let $b'_a(T, g) \equiv \lceil (1-\delta) \frac{f(T^{(g)})}{\mathcal{K}_{\inf}(\nu_a, c_a \rho^\star)} \rceil$ if $a \in \mathcal{N}$ and let $b'_a(T, g) \equiv \lceil \frac{f(T^{(g)})}{(1-\delta)\mathcal{K}_{\inf}(\nu_a, c_a \rho_{a^\star})} \rceil$ if $a \in \mathcal{M}$. Similar arguments to those used to derive (21) in Section 6.1 show that, for T large enough so that $T^{(g)} \geq K$,

$$\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{ a \in \hat{\mathcal{A}}(t+1), \hat{\nu}_{a,N_a(t)} \in \mathcal{C}_{c_a \rho^\dagger, f(t)/N_a(t)} \right\}$$
$$\leq \sum_{n=1}^{T^{(g-1)}-K} \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{ \hat{\nu}_{a,n} \in \mathcal{C}_{c_a \rho^\dagger, f(T^{(g-1)})/n}, \tau_{a,n+1} = t+1 \right\}.$$

We split the sum over n into a sum S_1 from $n = 1$ to $b'_a(T, g)$ and a sum S_2 from $n = b'_a(T, g) + 1$ to $T^{(g-1)} - K$. For the latter sum, the fact that, for each n, $\tau_{a,n+1} = t + 1$ for at most one t in a given interval, yields that

$$S_2 \leq \sum_{n=b'_a(T,g)+1}^{T^{(g-1)}-K} \mathbb{P}\left\{ \hat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T^{(g-1)})/n} \right\}.$$

If $a \in \mathcal{N}$, then δ satisfying (32) yields that $b'_a(T, g) > \frac{f(T^{(g-1)})}{\mathcal{K}_{\inf}(\nu_a, c_a \rho_{a^\star})}$, and so the above sum is $o(\log T)$ by Lemma A.1. If $a \in \mathcal{M}$, then $b'_a(T, g) = \lceil \frac{f(T^{(g-1)})}{\mathcal{K}_{\inf}(\nu_a, c_a \rho_{a^\star})} \rceil$, and so again the above sum is $o(\log T)$.

We now bound S_1. Note that if $N_a(T^{(g)} - 1) > b'_a(T, g)$, then, for every $n \leq b'_a(T, g)$, $\tau_{a,n+1} < T^{(g)}$ and $S_1 = 0$ (the sum over t is void). Therefore,

$$S_1 \leq \sum_{n=1}^{b'_a(T,g)} \mathbb{P}\left\{ N_a(T^{(g)} - 1) \leq b'_a(T, g) \right\} = b'_a(T, g)\, \mathbb{P}\left\{ N_a(T^{(g)} - 1) \leq b'_a(T, g) \right\}.$$

From Lemma 20, KL-UCB is uniformly efficient. Thus, by (8), for any $a \in \mathcal{M}$ one has

$$\lim_{T \to \infty} \mathbb{P}\left(N_a(T^{(g)} - 1) \leq \frac{2 \log(T^{(g)})}{(1-\delta)\, \mathcal{K}_{\inf}(\nu_a, c_a \rho_{a^\star})} \right) = 0,$$

where we use the fact that $\mathcal{K}_{\inf}(\nu_a, c_a \rho^\star) = 0$ and choose $\epsilon = (1-\delta)\, \mathcal{K}_{\inf}(\nu_a, c_a \rho_{a^\star})/2 > 0$. This yields that $\mathbb{P}\left\{ N_a\left(T^{(g)} - 1\right) < b'_a(T, g) \right\} \to 0$ as $T \to \infty$ and $S_1 = o(\log T)$.

If $a \in \mathcal{N}$, then Lemma 20 and (8) from Lemma 4 yield that

$$\mathbb{P}\left\{ N_a\left(T^{(g)} - 1\right) < (1-\delta) \frac{\log T^{(g)}}{\mathcal{K}_{\inf}(\nu_a, c_a \rho^\star)} \right\} \to 0 \text{ as } T \to 0.$$

The fact that $\lim_T f(T)/\log T = 1$ shows that $b'_a(T, g) = (1-\delta) \frac{\log T^{(g)}}{\mathcal{K}_{\inf}(\nu_a, c_a \rho^\star)} + o(\log T)$. Plugging this into (8) from Lemma 4 (which holds for *every* δ between 0 and 1) yields that $\mathbb{P}\left\{ N_a\left(T^{(g)} - 1\right) < b'_a(T, g) \right\} \to 0$ as $T \to \infty$. It follows that $S_1 = o(\log T)$.

We have then shown that $\mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})] = o(\log T)$ for each $a \in \underline{\mathcal{U}} \subset \mathcal{M} \cup \mathcal{N}$ and each $g \leq G$. As Term B is a sum of finitely many such terms, Term B is $o(\log T)$. □ □

D Supplementary Proofs for Thompson Sampling

We begin with a lemma.

Lemma A.3. *For any fixed real number L, arm a, $\mu_a < \mu^\dagger < \theta^\dagger$, and $t \geq 1$,*

$$I\left\{\hat{\mu}_a(t) \leq \mu^\dagger, N_a(t) \geq L\right\} \mathbb{P}\left\{\theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right\} \leq e^{-(L+1)\,\mathrm{KL}(\mu^\dagger, \theta^\dagger)}.$$

Proof. From Fact 3 in Agrawal and Goyal [2012] [also used in Agrawal and Goyal, 2011, Kaufmann et al., 2012a,b],

$$\mathbb{P}\left(\theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right) = \mathbb{P}\left(\sum_{n=1}^{N_a(T)+1} Z_n \leq \sum_{n=1}^{N_a(T)} \mathbb{1}\{X_{a,n} = 0\} \middle| \mathcal{F}(t)\right),$$

where $\{Z_n\}$ is an i.i.d. sequence (independent of all other quantities under consideration) of Bernoulli random variables with mean $\theta^\dagger$. Upper bounding the right-hand side yields

$$\mathbb{P}\left(\theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right) \leq \mathbb{P}\left(\frac{1}{N_a(T)+1} \sum_{n=1}^{N_a(T)+1} Z_n \leq \hat{\mu}_a(T) \middle| \mathcal{F}(t)\right).$$

Using that $\mu^\dagger < \theta^\dagger$, the Chernoff-Hoeffding bound gives that $\mathbb{P}\left(\theta_a(t) > \theta^\dagger \middle| \mathcal{F}(t)\right)$ is no larger than $e^{-[N_a(t)+1]\,\mathrm{KL}(\hat{\mu}_a(t), \theta^\dagger)}$. Multiplying the left-hand side by $I\left\{\hat{\mu}_a(t) \leq \mu^\dagger, N_a(t) \geq L\right\}$, this yields the upper bound $e^{-(L+1)\,\mathrm{KL}(\mu^\dagger, \theta^\dagger)}$. □ □

Proof of Lemma 11. Let $\tilde{\theta}_{\tilde{a}}(t) = \theta_{\tilde{a}}(t)$ for all $\tilde{a} \neq a^\star$ and let $\tilde{\theta}_{a^\star}(t) = -\infty$. Define the event $B \equiv \left\{\rho^\star(\tilde{\theta}_a(t) : a = 1, \ldots, K+1) < \theta_a(t)/c_a \leq \rho^\ddagger\right\}$. Observe that

$$\begin{aligned}
&\mathbb{P}\left\{\hat{\rho}^\star(t) \leq \frac{\theta_a(t)}{c_a} \leq \rho^\ddagger, \frac{\theta_{a^\star}(t)}{c_{a^\star}} < \hat{\rho}^\star(t) \middle| \mathcal{F}(t)\right\} \\
&= \mathbb{P}\left(\left\{\hat{\rho}^\star(t) \leq \frac{\theta_a(t)}{c_a}, \frac{\theta_{a^\star}(t)}{c_{a^\star}} < \hat{\rho}^\star(t)\right\} \cap B \middle| \mathcal{F}(t)\right) \\
&\leq \mathbb{P}\left(\left\{\frac{\theta_{a^\star}(t)}{c_{a^\star}} \leq \rho^\ddagger\right\} \cap B \middle| \mathcal{F}(t)\right). \qquad \text{(A.8)}
\end{aligned}$$

The event $\{\theta_{a^\star}(t)/c_{a^\star} > \rho^\ddagger\}$ is independent of the event B conditional on $\mathcal{F}(t)$, and so the fact that $\left\{\theta_{a^\star}(t)/c_{a^\star} > \rho^\ddagger\right\} \cap B \subseteq \{\theta_{a^\star}(t)/c_{a^\star} \geq \hat{\rho}^\star(t)\}$ yields

$$\mathbb{P}(B|\mathcal{F}(t)) \leq \frac{\mathbb{P}\left(\theta_{a^\star}(t)/c_{a^\star} \geq \hat{\rho}^\star(t) \middle| \mathcal{F}(t)\right)}{\mathbb{P}\left(\theta_{a^\star}(t)/c_{a^\star} > \rho^\ddagger \middle| \mathcal{F}(t)\right)}.$$

We note that $\mathbb{P}\left(\theta_{a^\star}(t) > c_{a^\star}\rho^\ddagger \middle| \mathcal{F}(t)\right)$ is positive (a beta distribution with at least one success is larger than $c_{a^\star}\rho^\ddagger < 1$ with positive probability). Finally, since $a \in \hat{\mathcal{A}}(t+1)$ implies that $\theta_{a^\star}(t)/c_{a^\star} \geq \hat{\rho}^\star(t)$, (A.8) yields

$$\begin{aligned}
&\mathbb{P}\left(a \in \hat{\mathcal{A}}(t+1), \theta_a(t)/c_a \leq \rho^\ddagger, \theta_{a^\star}(t)/c_{a^\star} < \hat{\rho}^\star(t) \middle| \mathcal{F}(t)\right) \\
&\leq \mathbb{P}\left(\left\{\theta_{a^\star}(t)/c_{a^\star} \leq \rho^\ddagger\right\} \cap B \middle| \mathcal{F}(t)\right) \\
&= \mathbb{P}\left(\theta_{a^\star}(t)/c_{a^\star} \leq \rho^\ddagger \middle| \mathcal{F}(t)\right) \mathbb{P}\left(B \middle| \mathcal{F}(t)\right) \\
&\leq \mathbb{P}\left(\theta_{a^\star}(t)/c_{a^\star} \leq \rho^\ddagger \middle| \mathcal{F}(t)\right) \frac{\mathbb{P}\left(\theta_{a^\star}(t)/c_{a^\star} \geq \hat{\rho}^\star(t) \middle| \mathcal{F}(t)\right)}{\mathbb{P}\left(\theta_{a^\star}(t)/c_{a^\star} > \rho^\ddagger \middle| \mathcal{F}(t)\right)}.
\end{aligned}$$

□ □

Proof of Lemma 12. Using (24), one can write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{1-p^{\rho^\ddagger}_{a^\star}(t)}{p^{\rho^\ddagger}_{a^\star}(t)}\,\mathbb{P}\left(\frac{\theta_{a^\star}(t)}{c_{a^\star}}\geq\hat{\rho}^\star(t)\middle|\mathcal{F}(t)\right)\right]\\
&\leq \mathring{q}_{a^\star}^{-1}\,\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{1-p^{\rho^\ddagger}_{a^\star}(t)}{p^{\rho^\ddagger}_{a^\star}(t)}\,\mathbb{P}\left(a^\star\in\hat{\mathcal{A}}(t+1)\middle|\mathcal{F}(t)\right)\right]\\
&= \mathring{q}_{a^\star}^{-1}\,\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{1-p_{a^\star,N_{a^\star}(t)}}{p_{a^\star,N_{a^\star}(t)}}\,\mathbb{1}\left\{a^\star\in\hat{\mathcal{A}}(t+1)\right\}\right]\\
&= \mathring{q}_{a^\star}^{-1}\,\mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{n=0}^{T-1}\frac{1-p^{\rho^\ddagger}_{a^\star,n}}{p^{\rho^\ddagger}_{a^\star,n}}\,\mathbb{1}\left\{\tau_{a^\star,n+1}=t+1\right\}\right]\\
&\leq \mathring{q}_{a^\star}^{-1}\,\mathbb{E}\left[\sum_{n=0}^{T-1}\frac{1-p^{\rho^\ddagger}_{a^\star,n}}{p^{\rho^\ddagger}_{a^\star,n}}\right],
\end{aligned}$$

where the latter inequality holds because $\tau_{a^\star,n+1}=t+1$ for at most one t in $\{0,\dots,T-1\}$. □ □

Proof of Lemma 14. Let $L^\dagger(T)\equiv\frac{\log T}{\mathrm{KL}(c_a\rho^\dagger,c_a\rho^\ddagger)}$. We have that

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{P}\left\{a\in\hat{\mathcal{A}}(t+1),\theta_a(t)>c_a\rho^\ddagger,\hat{\mu}_a(t)\leq c_a\rho^\dagger\right\}\\
&\quad=\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{1}\left\{N_a(t)<L^\dagger(T)-1,\hat{\mu}_a(t)\leq c_a\rho^\dagger\right\}\mathbb{P}\left(a\in\hat{\mathcal{A}}(t+1),\theta_a(t)>c_a\rho^\ddagger\middle|\,\mathcal{F}(t)\right)\right]\\
&\qquad+\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{1}\left\{N_a(t)\geq L^\dagger(T)-1,\hat{\mu}_a(t)\leq c_a\rho^\dagger\right\}\mathbb{P}\left(a\in\hat{\mathcal{A}}(t+1),\theta_a(t)>c_a\rho^\ddagger\middle|\,\mathcal{F}(t)\right)\right]\\
&\quad\leq\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{1}\left\{N_a(t)<L^\dagger(T)-1\right\}\mathbb{P}\left(a\in\hat{\mathcal{A}}(t+1)\middle|\,\mathcal{F}(t)\right)\right]\\
&\qquad+\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{1}\left\{N_a(t)\geq L^\dagger(T)-1,\hat{\mu}_a(t)\leq c_a\rho^\dagger\right\}\mathbb{P}\left(\theta_a(t)>c_a\rho^\ddagger\middle|\,\mathcal{F}(t)\right)\right].
\end{aligned}\tag{A.9}$$

The first term in the right hand side equals $\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbb{1}\left\{N_a(t)<L^\dagger(T)-1,a\in\hat{\mathcal{A}}(t+1)\right\}\right]$. Hence it is no larger than $L^\dagger(T)-1$ (the sum has at most $L^\dagger(T)-1$ nonzero terms). For the second term, Lemma A.3 yields

$$\mathbb{1}\left\{\hat{\mu}_a(t)<c_a\rho^\dagger,N_a(T)\geq L^\dagger(T)-1\right\}\mathbb{P}\left(\theta_a(t)>c_a\rho^\ddagger\middle|\,\mathcal{F}(t)\right)\leq e^{-L^\dagger(T)\,\mathrm{KL}(c_a\rho^\dagger,c_a\rho^\ddagger)}=T^{-1}.$$

It follows that the second term on the right of (A.9) is upper bounded by $\sum_{t=0}^{T-1}T^{-1}=1$. This completes the proof. □ □

Proof of Lemma 18 for Thompson sampling in the setting of Theorem 8. Fix $a\in\underline{\mathcal{U}}\cup\overline{\mathcal{U}}$ and $\epsilon\in(0,1)$. For ease of notation, we analyze $\mathbb{E}[M_a^{a^\star}(T)]$ rather than $\mathbb{E}[M_a^{a^\star}(T^{(G)})]$, but for fixed $G<\infty$ there is no loss of generality in doing so. If $a\in\underline{\mathcal{U}}$, then let $\mu^\dagger=c_a\rho_{a^\star}$, and otherwise, fix $\mu^\dagger\in(\mu_a,\mu_+)$. Let $\rho^\dagger$ and $\rho^\ddagger$ satisfy $\rho_a<\rho^\dagger<\rho^\ddagger<\mu^\dagger/c_a$ (exact quantities to be specified at the end of the proof). Note that

$$\left\{a\in\hat{\mathcal{A}}(t+1),\frac{\theta_{a^\star}(t)}{c_{a^\star}}<\hat{\rho}^\star(t)\right\}$$

$$\subseteq \left\{a \in \hat{\mathcal{A}}(t+1), \frac{\theta_a(t)}{c_a} \le \rho^\ddagger, \frac{\theta_{a^\star}(t)}{c_{a^\star}} < \hat{\rho}^\star(t)\right\} \cup \left\{a \in \hat{\mathcal{A}}(t+1), \frac{\theta_a(t)}{c_a} > \rho^\ddagger\right\}.$$

Recalling that $\mathbb{E}[M_a^{a^\star}(T)]$ is equal to $\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \theta_{a^\star}(t)/c_{a^\star} < \hat{\rho}^\star(t)\right\}$, the above yields

$$\begin{aligned}\mathbb{E}[M_a^{a^\star}(T)] \le &\sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \frac{\theta_a(t)}{c_a} \le \rho^\ddagger, \frac{\theta_{a^\star}(t)}{c_{a^\star}} < \hat{\rho}^\star(t)\right\} \\ &+ \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \frac{\hat{\mu}_a(t)}{c_a} > \rho^\dagger\right\} \\ &+ \sum_{t=0}^{T-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \frac{\theta_a(t)}{c_a} > \rho^\ddagger, \frac{\hat{\mu}_a(t)}{c_a} \le \rho^\dagger\right\}. \end{aligned} \tag{A.10}$$

Note that the right-hand side of the above is almost identical to (23). Note that all of the results used to control the three terms on the right-hand side of (23) hold for any a with $\rho_a \le \rho^\star$ provided $\rho_a < \rho^\dagger < \rho^\ddagger < \mu^\dagger/c_a$. In particular, we are referring to Lemma 11, (25), Lemma 13, (26), Lemma 15, and Lemma 14. Hence, $\mathbb{E}[M_a^{a^\star}(T)] \le \frac{\log T}{\mathrm{KL}(c_a\rho^\dagger, c_a\rho^\ddagger)} + o(\log T)$.

Selecting $\rho^\dagger$ and $\rho^\ddagger$ as in the proof of (11) and (12) from Theorem 8 yields $\mathbb{E}[M_a^{a^\star}(T)] \le (1+\epsilon)^2 \frac{\log T}{\mathrm{KL}(\mu_a, \mu^\dagger)} + o(\log T)$. As ϵ was arbitrary, dividing both sides by $\log T$ and taking $T \to \infty$ followed by $\epsilon \to 0$ yields that $\mathbb{E}[M_a^{a^\star}(T)] \le \frac{\log T}{\mathrm{KL}(\mu_a, \mu^\dagger)} + o(\log T)$. If $a \in \underline{\mathcal{U}}$, then replacing T by $T^{(G)}$ (for T large enough so that $T^{(G)} > 1$) gives the desired $\mathbb{E}[M_a^{a^\star}(T^{(G)})] \le (1-\delta)^G \frac{\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\dagger)} + o(\log T)$ in light of the fact that $\mu^\dagger = c_a\rho_{a^\star}$. If, on the other hand, $a \in \overline{\mathcal{U}}$, then the same arguments used to conclude the $a \in \overline{\mathcal{U}}$ result in the proof of Lemma 18 for KL-UCB, namely selecting an appropriate sequence $\mu^\dagger(T) \to \mu_+$, can be used to show that $\mathbb{E}[M_a^{a^\star}(T^{(G)})] = o(\log T)$. □ □

Proof of Lemma 19 for Thompson sampling in the setting of Theorem 8. Fix $g \in \mathbb{N}$, an arm $a \in \underline{\mathcal{U}} \subset \mathcal{M} \cup \mathcal{N}$, and $T^{(g)}$ such that $T^{(g)} > 1$. Let $\rho^\dagger$ and $\rho^\ddagger$ satisfy $\rho_a < \rho^\dagger < \rho^\ddagger < \rho_{a^\star}$ and $\mathrm{KL}(c_a\rho^\dagger, c_a\rho^\ddagger) \ge (1-\delta)\,\mathrm{KL}(\mu_a, c_a\rho_{a^\star})$. By the same arguments used for (A.10),

$$\begin{aligned}&\mathbb{E}[M_a^{a^\star}(T^{(g-1)}) - M_a^{a^\star}(T^{(g)})] \\ &\le \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \frac{\theta_a(t)}{c_a} \le \rho^\ddagger, \frac{\theta_{a^\star}(t)}{c_{a^\star}} < \hat{\rho}^\star(t)\right\} \\ &\quad + \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \frac{\hat{\mu}_a(t)}{c_a} > \rho^\dagger\right\} \\ &\quad + \sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \frac{\theta_a(t)}{c_a} > \rho^\ddagger, \frac{\hat{\mu}_a(t)}{c_a} \le \rho^\dagger\right\}. \end{aligned} \tag{A.11}$$

The first two sums are trivially upper bounded by the sums from $t = 0$ to $T-1$, and thus are $o(\log T)$ by Lemma 11, (25), Lemma 13, (26), and Lemma 15. If $a \in \mathcal{N}$, then let $b_a(T,g) \equiv (1-\delta)\frac{\log T^{(g)}}{\mathrm{KL}(\mu_a, c_a\rho^\star)}$, and if $a \in \mathcal{M}$ then let $b_a(T,g) \equiv \frac{\log T^{(g)}}{(1-\delta)\,\mathrm{KL}(\mu_a, c_a\rho_{a^\star})}$. We have that

$$\begin{aligned}&\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \frac{\theta_a(t)}{c_a} > \rho^\ddagger, \frac{\hat{\mu}_a(t)}{c_a} \le \rho^\dagger\right\} \\ &= \mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1} \mathbb{1}\left\{\frac{\hat{\mu}_a(t)}{c_a} \le \rho^\dagger, N_a(t) \ge b_a(T,g)\right\} \mathbb{P}\left\{a \in \hat{\mathcal{A}}(t+1), \frac{\theta_a(t)}{c_a} > \rho^\ddagger \middle| \mathcal{F}(t)\right\}\right] \end{aligned}$$

$$+\mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1}\mathbb{1}\left\{\frac{\hat{\mu}_a(t)}{c_a}\le\rho^\dagger, N_a(t)<b_a(T,g)\right\}\mathbb{P}\left\{a\in\hat{\mathcal{A}}(t+1),\frac{\theta_a(t)}{c_a}>\rho^\ddagger\middle|\mathcal{F}(t)\right\}\right]. \tag{A.12}$$

If $a\in\mathcal{N}$, then Lemma A.3 and $\mathrm{KL}(c_a\rho^\dagger, c_a\rho^\ddagger)\ge(1-\delta)\,\mathrm{KL}(\mu_a, c_a\rho_{a^\star})$ yield that the first term on the right is upper bounded by

$$\begin{aligned}&\sum_{t=T^{(g)}}^{T^{(g-1)}-1}\exp\left[-(1-\delta)^2\frac{\log T^{(g-1)}}{\mathrm{KL}(\mu_a,c_a\rho^\star)}\,\mathrm{KL}(\mu_a,c_a\rho_{a^\star})\right]\\&\qquad\le T^{(g-1)}\exp\left[-(1-\delta)^2\frac{\log T^{(g-1)}}{\mathrm{KL}(\mu_a,c_a\rho^\star)}\,\mathrm{KL}(\mu_a,c_a\rho_{a^\star})\right]\le 1,\end{aligned}$$

where the second inequality holds because δ satisfies (32). If $a\in\mathcal{M}$, then we instead have that this term is no larger than

$$\sum_{t=T^{(g)}}^{T^{(g-1)}-1}\exp\left[-\frac{\log T^{(g-1)}}{\mathrm{KL}(\mu_a,c_a\rho_{a^\star})}\,\mathrm{KL}(\mu_a,c_a\rho_{a^\star})\right]\le 1.$$

For the second term in (A.12), note that

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1}\mathbb{1}\left\{\frac{\hat{\mu}_a(t)}{c_a}\le\rho^\dagger, N_a(t)<b_a(T,g)\right\}\mathbb{P}\left\{a\in\hat{\mathcal{A}}(t+1),\frac{\theta_a(t)}{c_a}>\rho^\ddagger\middle|\mathcal{F}(t)\right\}\right]\\&\le\mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1}\mathbb{1}\left\{N_a(t)<b_a(T,g)\right\}\mathbb{P}\left\{a\in\hat{\mathcal{A}}(t+1)\middle|\mathcal{F}(t)\right\}\right]\\&=\mathbb{E}\left[\sum_{t=T^{(g)}}^{T^{(g-1)}-1}\mathbb{1}\left\{N_a(t)<b_a(T,g), a\in\hat{\mathcal{A}}(t+1)\right\}\right]\\&=\mathbb{E}\left[\mathbb{1}\left\{N_a\left(T^{(g)}\right)<b_a(T,g)\right\}\sum_{t=T^{(g)}}^{T^{(g-1)}-1}\mathbb{1}\left\{N_a(t)<b_a(T,g), a\in\hat{\mathcal{A}}(t+1)\right\}\right]\\&\le b_a(T,g)\,\mathbb{P}\left\{N_a\left(T^{(g)}\right)<b_a(T,g)\right\},\end{aligned}$$

where the final inequality uses that the sum inside the expectation is at most $b_a(T,g)$. By the uniform efficiency of the algorithm established in Lemma 20 and (8) from Lemma 4, the probability in the final inequality is $o(1)$, and thus the above is $o(b_a(T,g))=o(\log T)$. Thus (A.12) is $o(\log T)$.

Plugging this into (A.11) yields that $\mathbb{E}[M_a^{a^\star}(T^{(g-1)})-M_a^{a^\star}(T^{(g)})]=o(\log T)$ for each $a\in\underline{\mathcal{U}}\subset\mathcal{M}\cup\mathcal{N}$ and each $g\le G$. As Term B is a sum of finitely many such terms, Term B is $o(\log T)$. □ □